



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

Trend and reversal of idiosyncratic volatility revisited

Leippold, Markus ; Svaton, Michal

Abstract: We reexamine the existence of an upward trend in the idiosyncratic volatility (IV) of US-listed stocks between 1962 and 2000 documented in CLMX2001. Following the same methodology as in their paper, we confirm the upward trend and the subsequent reversal, as reported by BHZ2012. However, taking a closer look, we find that this result is influenced by microstructure biases in realized variances and correlations based on daily returns. Correcting these biases, we find that, depending on whether we use equal- or value-weighted IV and whether we include the pre-NASDAQ period, between 17

DOI: <https://doi.org/10.2139/ssrn.3321678>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-191300>

Journal Article

Accepted Version

Originally published at:

Leippold, Markus; Svaton, Michal (2019). Trend and reversal of idiosyncratic volatility revisited. Critical Finance Review:Epub ahead of print.

DOI: <https://doi.org/10.2139/ssrn.3321678>

Trend and Reversal of Idiosyncratic Volatility Revisited

Markus Leippold, *

Michal Svatoň, *

* University of Zurich, Department of Banking and Finance, Plattenstrasse 14, 8032 Zurich, Switzerland

ABSTRACT

We reexamine the existence of an upward trend in the idiosyncratic volatility (IV) of US-listed stocks between 1962 and 2000 documented in Campbell *et al.* (2001). Following the same methodology as in their paper, we confirm the upward trend and the subsequent reversal, as reported by Bekaert *et al.* (2012). However, taking a closer look, we find that this result is influenced by microstructure biases in realized variances and correlations based on daily returns. Correcting these biases, we find that, depending on whether we use equal- or value-weighted IV and whether we include the pre-NASDAQ period, between 17% and 62% of the trend is removed. Also, the subsequent reversal of IV is almost entirely removed once we correct for these biases. We also highlight the mechanical dependence of the IV measure on industry concentration. From 1962 to 2000, industry concentration also contributes to a higher IV. However, its effect remains quantitatively small.

We thank Catherine Bruneau, John Campbell, Thomas Conlon, John Cotter, Michel Dacorogna, Jean-Philippe Bouchaud, Andreas Höpner, Juhani Linnainmaa, Ivo Welch, Richard McGee, Giorgio Mirone, Per Ostberg, Viktor Todorov, Alexandre Ziegler, the semi-

Keywords: Idiosyncratic volatility, measurement error, bid-ask bounce, asynchronicity, CRSP database

JEL Codes: G11, G12, G14

1 Introduction

It is the consensus that the US equity market in the period from the 1960s to the early 2000s experienced a strengthening role of idiosyncratic (or firm-level) volatility (IV) relative to the systemic component of market risk, as discussed in Campbell *et al.* (2001), hereafter CLMX. Furthermore, recent studies indicate that the aggregate level of idiosyncratic risk reverted around 2000 to the levels observed in the 1960s and systemic risk regained its importance.¹ At the same time, there is a large body of literature that tries to explain these trends, but no general agreement has been reached about the underlying causes generating the observed trends.²

In this paper, we challenge the consensus. We first replicate the analysis of Campbell *et al.* (2001) and confirm their main result of the upward trend in idiosyncratic volatility in 1962-1997 period, as well as the reversal of the IV in more recent studies. However, we show that both of those empirical findings are, to a nontrivial degree, driven by the variation in the severity of microstructure biases of daily-return data in the CRSP database, and by

nar participants at the University of Zurich, the University College Dublin, the SwissQuant Scientific Board Meeting, the AFFI Conference 2018 (Paris), the 2018 Financial Econometrics Conference on Market Microstructure, Limit Order Books and Derivative Markets at the University of Lancaster, and the 9th International Conference of the Financial Engineering and Banking Society in Prague, and the anonymous referee or their feedback.

¹See, for example, Brown and Kapadia (2007), Bali *et al.* (2008), Brandt *et al.* (2010), and Bekaert *et al.* (2012). These studies provide various explanations of the change in IV, such as changes in institutional ownership, and growth options. Most recently, Bartram *et al.* (2018) argue that for 2013 to 2017 IV was historically low.

²A large literature tries to explain the trends in IV, for example, by an increase in competition (Comin and Philippon, 2005; Gaspar and Massa, 2006; Pontiff and Irvine, 2009), increased R&D spending and Schumpeterian destruction (Chun *et al.*, 2008; Comin and Mulani, 2009), financial development (Brown and Kapadia, 2007), institutional ownership (Xu and Malkiel, 2003; Bennett *et al.*, 2003), irrational exuberance (Brandt *et al.*, 2010), young firms (Fink *et al.*, 2010; Bekaert *et al.*, 2012), profitability (Wei and Zhang, 2006), and macroeconomic factors (Bekaert *et al.*, 2012). Bartram *et al.* (2018) use firm characteristics to explain why IV was historically low from 2013 to 2017, compared to the period from 1963 to 2012.

changes in industry concentration.³ When we account for microstructure biases, the trend is weaker by 17% in the value-weighted, and by 34% in the equal-weighted case. The contribution of microstructure biases is even larger when we consider a shorter window starting in 1972, when NASDAQ universe joined the CRSP database, with trend estimates lower by 44% in the value-weighted case, and by 62% in the equal-weighted case. Moreover, once we eliminate the biases, there is no longer a striking reversal in the IV series. The changes in the magnitude of the biases reflect events such as changes to tick size requirements, and data-related aspects, e.g., increasing availability of trading prices in the CRSP database.

To motivate our claim with an illustration, we Panel A of Figure 1 plots the price process for National Mobile Corp listed on the NASDAQ SmallCap. For this stock, the price quotation changed from midpoint to closing price around June 15, 1992. After this change, the price oscillates between bid and ask quotes, depending on whether the last order was a buy or a sell order. This phenomenon is commonly referred to as the bid-ask bounce. It results in higher volatility of closing prices compared to the underlying “efficient” price derived from the variation in expected (and appropriately discounted) future cash flows. Panel B of Figure 1 shows that the average effect of this switch on realized volatility across all stocks is roughly 100%. Hence, the realized volatility on average doubles when closing prices become available. Hence, when estimating IV, we must control for the type of price quotation, particularly for the early CRSP data where some price observations are quote midpoints rather than closing prices.⁴

The transition from midpoint quotations to closing prices has not been

³Changes in industry concentration correspond to the growth in the number of listed firms before 2000 and its recent decrease, which complicates the comparison of the CLMX measure of aggregate idiosyncratic volatility at different points in time.

⁴Because the bid-ask bounce results in less frequent zero returns, using the frequency of zero returns as a liquidity proxy, as proposed, e.g., by Lesmond *et al.* (1999), is problematic when using historical CRSP data, because it depends on the availability of closing price rather than liquidity. Microstructure effects also provide a plausible explanation that in the CLMX sample, if we move from daily to monthly sampling frequency, the average variance decreases, while average correlation and portfolio (industry, market) variances increase. Consistent with stronger microstructure effects for small, illiquid stocks, the choice of sampling frequency has a more pronounced effect on equally-weighted measures. Recent reversal coincides with exogenous events of decimalization. Both on market and stock level, all of those effects seem difficult to justify on the grounds of fundamental factors.

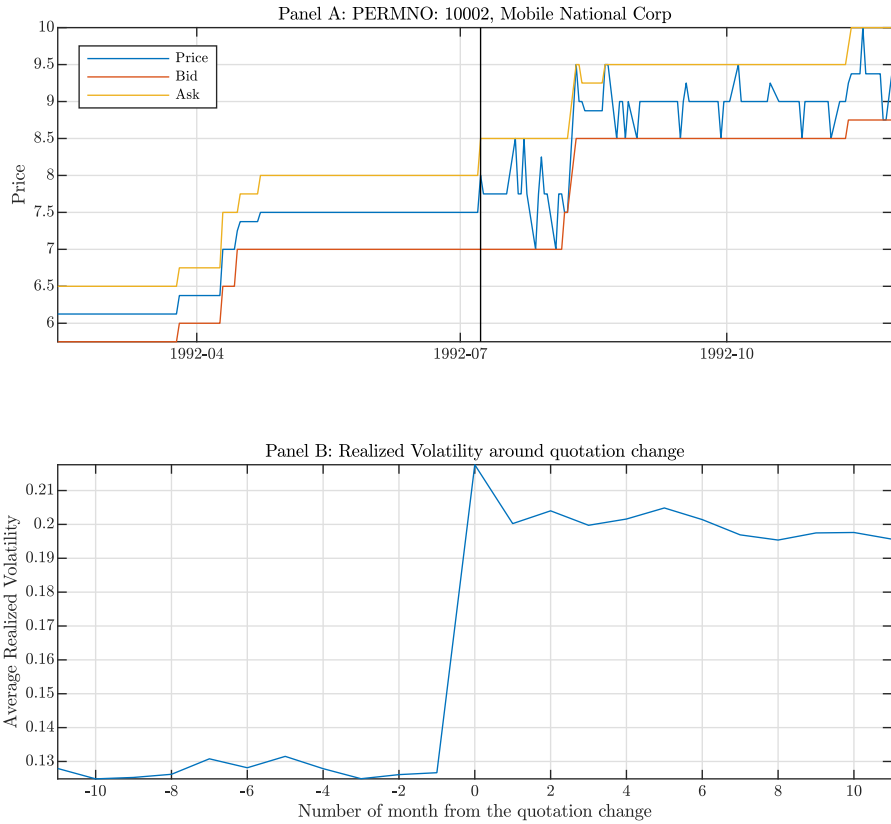


Figure 1: Midpoint vs. closing price realized volatility.

Description: Panel A: Example of the price series of a stock (National Mobile Corp) around NASDAQ SmallCap's change from midpoint quotation to closing price. The vertical line denotes the date of the change in price type. Panel B: Monthly realized volatility computed from daily returns of stocks around the first date of availability of their trading prices. We do not include stocks for which the first trading price is available in its first 60 observations because a lack of trading prices might reflect low liquidity rather than data availability. Month zero refers to the period of the first 22 days after the quotation change.

Interpretation: Stock prices in CRSP database are recorded either as a closing price or as a quote midpoint. Only closing prices are affected by bid-ask bounce, which inflates realized variances.

the only change. Between 1962 to 2001, a series of fundamental institutional changes occurred, all with the potential to substantially affect IV measurement.⁵ We classify them into three broad categories. The first category, as exemplified in Figure 1, relates to the way how prices are recorded in the CRSP database.⁶ The second category concerns the addition of new stocks to the CRSP database. For example, in December 1972, NASDAQ stocks were added, increasing the number of riskier firms in the sample. The third category relates to the tick size requirements. In 1997, the tick size reduction from $1/8$ to $1/16$ significantly reduced dollar bid-ask spreads on NYSE, NASDAQ, and Amex. Finally, in April 2001, all the stock exchanges adopted decimal quotes, which led to reduction of spreads. As a consequence, microstructure effects in daily returns quantitatively become negligible after this date.

The CLMX measure of idiosyncratic volatility depends on an interaction of volatilities, correlations, and weights assigned to individual stocks, which complicates the direct analysis of the role of microstructure effects and the role of changing the number of firms, or more generally, of capital concentration. To facilitate our empirical analysis, we provide an approximation of IV measure of CLMX as a product of average variance, average within-industry correlation, and industry concentration. With this approximation, we can match the IV series to a high degree of accuracy. Therefore, we analyze each of these channels separately. In particular, we ask which one is the main driver for the potential trend in the CLMX period and whether it is robust when we adjust for different microstructure biases.

We find that microstructure biases are mainly caused by the bid-ask bounce and price asynchronicity, both of which are amplified by price discreteness.⁷ Unsurprisingly, the bid-ask bounce impact is most substantial for stocks with wide effective spreads. However, it turns out that an addi-

⁵We give an overview of all these changes in Table 1 of the Online Appendix.

⁶For instance, for NASDAQ stocks, no trading prices were initially available, so the recorded prices were quote midpoints. As a consequence, realized variance estimates on NASDAQ in the 1970s were free of bid-ask bounce effects. For more information, we refer to the Data Description Guide for CRSP US Stock & US Index Databases (update as of March 31, 2017).

⁷Market microstructure typically entails a much broader set of phenomena, see Aït-Sahalia and Yu (2009) for a discussion. In this study, we use the term “market microstructure” to refer to effects that we control for in our analysis: bid-ask bounce, price discreteness, and asynchronicity.

tional critical factor is the way prices are recorded in the CRSP database because they can represent either a closing price or a quote midpoint. Trading prices, which are unlike quote midpoints prone to the bid-ask bounce, were unavailable for any NASDAQ stocks before 1982 and for NASDAQ SmallCap stocks before 1992. This subtle difference leads to substantial bias in variances only in the 1980s and not immediately after the inclusion of NASDAQ in the sample.

In addition to the bid-ask bounce, asynchronicity and price discreteness further impact correlation estimates. Indeed, these effects explain the observation of CLMX that, in their sample, the correlation estimates based on daily frequency are well below those obtained with monthly returns. Moreover, the bias in correlations reconciles the fact that in the CLMX sample, industry and market variances are higher on lower frequencies, despite the abundant evidence of bid-ask bounce effect, which is weaker at lower frequencies. The data also show other symptoms of the asynchronicity such as the incidence of zero daily returns, which peaks in the CLMX sample at 50% of daily returns in a given month. After quote decimalization around 2001, the wedge between correlation estimates based on different frequencies disappears, and so does the frequency of zero returns. In turn, the shrinking of the asynchronicity bias contributes to a drop in the IV estimates. For the value-weighted IV, the effect of correlation biases is modest because most weight lies on correlations between large, liquid stocks. However, the channel remains important for equal-weighted measure and for industry and market components.

Having identified the primary sources of microstructure biases, we explicitly correct for the biases in two ways. First, for the variances, we not only borrow from the literature (Roll, 1984; Hasbrouck, 2009), but we also develop a novel model, which we tailor to take into account features of both the microstructure and the CRSP data. Second, for the correlations, we provide a simple approximation of the bias based on the frequency of zero returns. The bias-corrected results show that the average variances on individual exchanges are free of a trend. On the aggregate market, a value-weighted measure of IV exhibits a stable long-term mean, only with surprisingly low variance in the very beginning of the sample period. The equal-weighted IV retains a small portion of the trend, which reflects the fact that NASDAQ, composed of stocks with higher variance, grew faster in numbers than NYSE in that period.

We also explore the dependence of the CLMX measure of IV on industry concentration, which establishes a direct link with market size and the industry classification used.⁸ Empirically, the industry concentration channel explains only a part of the IV changes and has a stronger influence on the systemic risk components. Its effect is stronger for value-weighted IV (roughly 18% of the trend estimate), for which the concentration varied more substantially.

In a related study, Lesmond *et al.* (2018), hereafter LPSZ, attribute the trend to the bid-ask bounce, which was eliminated with quote decimalization.⁹ We deviate from LPSZ along three dimensions. First, we show that the trend is not driven purely by bias in variance estimates, but also by microstructure biases in correlation, and by variation in industry concentration. Second, we introduce a novel model that explicitly corrects for microstructure biases. The model also accounts for the structural changes (e.g., tick-size reductions) and data-related events (e.g., changes to the availability of quote and price data), which is crucial for explaining the timing of the IV increase observed by CLMX.¹⁰ Third, we consider not only the value-weighted IV but also the equal-weighted measure. It provides a more robust “test” of the microstructure explanation because it puts more substantial weight on small stocks, which are more prone to microstructure effects.

We organize the paper as follows. In Section 2, we review the CLMX framework for IV measurement and provide its accurate approximation, which decomposes the IV to three separate drivers. In Section 3, we analyze the individual drivers and the role of microstructure biases in their

⁸Other industry classification schemes, for example, the text-based industry classification of Hoberg and Phillips (2010) and Hoberg and Phillips (2017), might be superior to the one employed in CLMX, but typically do not cover the period under consideration.

⁹However, the regression approach of LPSZ, which is the only part of their analysis covering the entire CLMX sample, raises some concerns about endogeneity, as we argue in our Online Appendix, Section B.

¹⁰As illustrated in Panel B of Figure 1, the different types of price quotations play a crucial role in the effect of the bid-ask bounce on IV estimation. The regression-based approach of LPSZ cannot account for these differences. However, they are essential for understanding the timing of the IV increase in the pre-decimalization period. They become especially relevant for the NASDAQ stock universe, for which the IV increased despite flat or even decreasing spread, which is in contradiction to the pure spread-bias relation studied by LPSZ.

measurement. We discuss their effect on the IV in Section 4. Section 5 concludes.

2 Revisiting CLMX

As a starting point, we review the approach of CLMX to decompose stock returns into three components: a market return, an industry-specific return, and a firm-specific return. Following their notation, we denote the excess market return at time t by R_{mt} , the excess return of industry i by R_{it} , and the excess return of firm j belonging to industry i by R_{jit} . For the market return, CLMX assume that the capital asset pricing model holds. Hence, we can write industry and firm returns as

$$R_{it} = \beta_{im}R_{mt} + \tilde{\epsilon}_{it}, \quad (1)$$

$$R_{jit} = \beta_{ji}R_{it} + \tilde{\eta}_{jit} = \beta_{ji}\beta_{im}R_{mt} + \beta_{ji}\tilde{\epsilon}_{it} + \tilde{\eta}_{jit}, \quad (2)$$

where $\tilde{\epsilon}_{it}$ and $\tilde{\eta}_{jit}$ denote the industry and firm innovations, and β_{im} and β_{ji} are the corresponding beta coefficients. To avoid the estimation of market betas, CLMX follow Campbell *et al.* (1997, Chapter 4) and use the “market-adjusted-return model” with the following modified residuals:

$$\epsilon_{it} = R_{it} - R_{mt}, \quad \eta_{jit} = R_{jit} - R_{it}. \quad (3)$$

By w_{it} we denote the weight of industry i in the total market and by w_{jit} the weight of firm j in industry i . Then,

$$\sigma_{\epsilon t}^2 := \sum_i w_{it} \sigma_{\epsilon it}^2, \quad (4)$$

$$\sigma_{\eta t}^2 := \sum_i w_{it} \sum_{j \in i} w_{jit} \sigma_{\eta jit}^2, \quad (5)$$

are the variance of the industry and firm-level residuals, respectively. We refer to the variance of the firm-level residuals interchangeably as the IV. We follow CLMX and estimate the individual variances $\sigma_{\eta jit}^2$ ($\sigma_{\epsilon it}^2$) using the realized variance, summing the squared daily returns of a firm (industry) within a given month. For the market model, the terms in Equations (4) and (5) are defined analogously, to obtain $\sigma_{\epsilon t}^2$ and $\sigma_{\eta t}^2$. The estimates

of variances for the market model are related to the variances for the market-adjusted model by

$$\sigma_{\epsilon t}^2 = \sigma_{\tilde{\epsilon} t}^2 + \text{CSV}_t(\beta_{im})\sigma_{mt}^2, \quad (6)$$

$$\sigma_{\eta t}^2 = \sigma_{\tilde{\eta} t}^2 + \text{CSV}_t(\beta_{im})\sigma_{mt}^2 + \text{CSV}_t(\beta_{ji})\sigma_{\tilde{\epsilon} t}^2, \quad (7)$$

where σ_{mt} is the market volatility and

$$\text{CSV}_t(\beta_{im}) := \sum_i w_{it}(\beta_{im} - 1)^2, \quad \text{CSV}_t(\beta_{ji}) := \sum_i w_{it} \sum_{j \in i} w_{jit}(\beta_{ji} - 1)^2$$

denote the cross-sectional variations of the betas, which drive the deviations of the actual average variances from those of the model and from those of the “market-adjusted” model. The advantage of using the market-adjusted model is that it avoids the explicit estimation of the market betas.¹¹

To be confident that we have built our analysis on the same subset of data as CLMX, we revisit their main empirical results. We use stock price data from CRSP spanning the period from January 1962 to December 2016. Hence, compared to the study of CLMX, which spans the period from July 1962 to December 1997, our sample period starts six months earlier and ends nineteen years later. As a consequence, our dataset includes a jump in the number of firms in July 1962, when Amex stocks were added to the database. A second spike occurred in December 1972, with the inclusion of NASDAQ stocks. For our analysis, we include all common stocks listed on NYSE, Amex, and NASDAQ with share codes 11 and 12.¹² As in Campbell *et al.* (2001), we follow Fama and French (1997) and aggregate individual firms into 49 industries according to their SIC classification.

Figure 2 shows the estimates of individual variance components. For reference, we include the series constructed by CLMX. Visually, the replication is almost perfect, with the market and firm components of CLMX

¹¹Its use is justified by the small magnitude of the cross-sectional variations of the betas. In particular, we find that the correlation between the increments of the exact and the “market-adjusted” computation is 0.89 and 0.97 for industry and firm variances, respectively. Hence, we find the quality of the approximation to be adequate, especially for the IV, our main object of interest.

¹²This restriction is standard, and its impact on the final result is negligible. Moreover, the notion of IV for funds (e.g., index ETFs) has a very different interpretation from that for common stocks.

mostly hidden behind the lines corresponding to our replication of the components mentioned above. The only significant deviation is in the industry component in the 1975-1985 period. One possible explanation of this gap is that we use a more recent SIC grouping version, published on Kenneth French's website. However, the gap does not disappear even when we use the old version of SIC code grouping from Fama and French (1997). The same applies to other changes in the construction of variance components we tested, such as exclusions of secondary share classes, or rebalancing the portfolios daily according to their capitalization instead of using the initial one. Panel A also confirms the conclusion of more recent studies (e.g., Bekaert *et al.* (2012)) that after the CLMX period, the IV reverted to the levels observed in the 1960s, while industry and market components appear higher in post-CLMX period than before.

While Figure 2 gives a visual inspection about how well our data matches the data in CLMX, Table 1 reports the numbers related to our replication of the CLMX results. The results show that our replication is accurate, and it confirms the main conclusions of CLMX: Idiosyncratic volatility exhibits a significant trend. In contrast, for other components (industry and market), the trend is often insignificant. As in CLMX, we perform an additional robustness check and replace the largest value (October 1987) by the second largest in the sample and report the columns labeled "Downweighted crash". This additional exercise shows that a single outlier does not drive the results. In the Online Appendix C (Tables 2 to 4), we provide additional replication results for the equal-weighted (EW) IV, and for the IV series based on lower return frequencies. We find the same patterns as CLMX. We confirm that, in their setup, the idiosyncratic volatility trend is stronger for the EW measure and weaker at lower frequencies.

Nevertheless, compared to the initial results in CLMX, some small numerical deviations remain. For instance, we estimate a trend of 1.03×10^{-5} for the idiosyncratic volatility, compared to a trend of 0.965×10^{-5} in CLMX. These numerical deviations likely stem from a combination of differences in inputs, i.e., the estimated volatility components, from how we and CLMX compute the statistics reported in the table, or from numerical instabilities because the trend is a tiny number. Figure 2 shows that the largest difference in the trend test input, i.e., the variance component series, is in the industry component. To investigate whether this input difference is the cause of numerical differences in trend estimates, we reverse-engineered

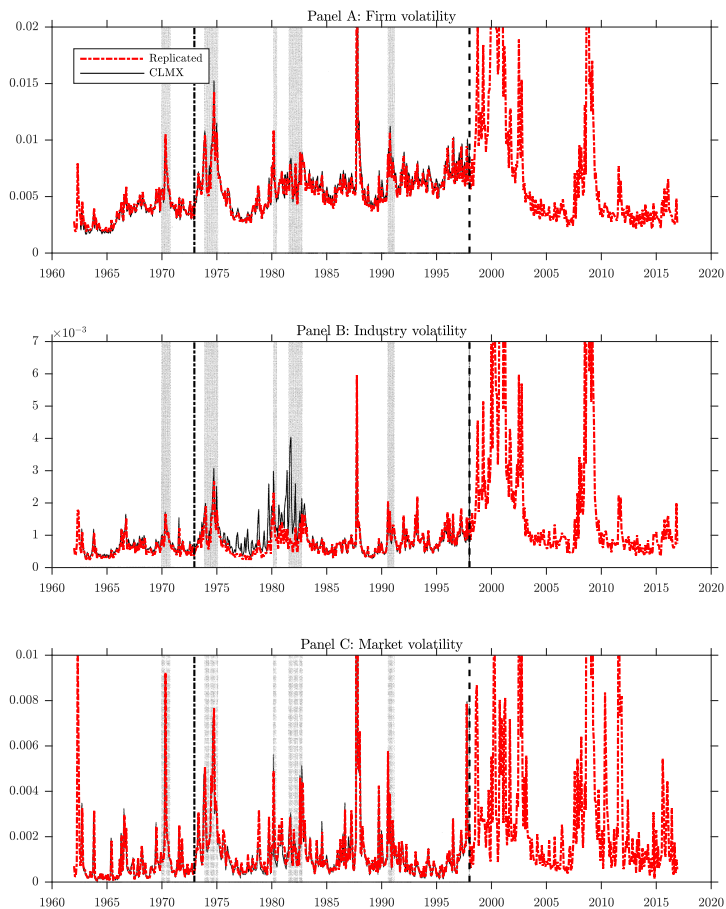


Figure 2: CLMX replication.

Description: CLMX variance decomposition of total variance into idiosyncratic (firm, Panel A), industry (Panel B), and market (Panel C) components. For reference, we include the original graph of the times series displayed in Campbell *et al.* (2001) as background for our figure. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: The replication of CLMX series is visually almost perfect, with the only visible deviation in the industry component in 1975-1985. The idiosyncratic volatility trends up in the CLMX sample and reverts afterwards.

	Raw data			Downweighted crash		
	MKT	IND	FRM	MKT	IND	FRM
Daily, VW (replicated), (N = 426)						
Mean × 100	1.56	0.90	6.34	1.43	0.89	6.28
Std. dev. × 100	3.45	0.54	2.75	1.48	0.46	2.28
Std. dev. × 100 detrended	3.44	0.53	2.44	1.47	0.45	1.93
Linear trend × 10 ⁵	0.20	0.11	1.03	0.12	0.11	1.00
PS-statistic	0.28	0.08	0.88	0.16	0.07	0.83
Confidence interval	(0.00, 0.55)	(0.03, 0.13)	(0.54, 1.21)	(-0.08, 0.40)	(0.02, 0.12)	(0.48, 1.18)
Daily, VW (original), (N = 426)						
Mean × 100	1.54	1.03	6.44	1.41	1.03	6.38
Std. dev. × 100	3.50	0.66	2.91	1.47	0.62	2.45
Std. dev. × 100 detrended	3.49	0.66	2.54	1.46	0.62	2.01
Linear trend × 10 ⁵	0.16	0.06	0.96	0.09	0.06	0.94
PS-statistic	0.26	0.09	1.00	0.14	0.08	0.96
Confidence interval	(-0.07, 0.60)	(-0.10, 0.27)	(0.55, 1.47)	(-0.12, 0.41)	(-0.10, 0.27)	(0.49, 1.42)

Table 1: Replication of main CLMX results.

Description: Replication of Campbell *et al.* (2001, Table III) on the CLMX sample (July 1962 - December 1997) for value-weighted IV computed with daily returns. The means and standard deviations are annualized. The standard deviation detrended correspond to a standard deviation of residuals in a regression of variance component on linear trend. Linear trend statistic is based on OLS of the variance measure on a linear trend. PS statistic is computed as in Vogelsang (1998); the corresponding implied confidence interval is computed on 90% level. Panel labeled “replicated” shows the values computed by us, panel labeled “original” shows the corresponding numbers as presented in Campbell *et al.* (2001). Raw data columns compute the statistics directly from estimates of individual variance components. For downweighted crash columns, we replace the largest observation by the second largest value over the CLMX (1962-1997) sample.

Interpretation: Our replication is reasonably accurate and confirms the main conclusion of Campbell *et al.* (2001): Idiosyncratic volatility trended up in period 1962-1997, while the other components did not.

the volatility components' data points from Figures 2 to 4 in CLMX by converting their original graph to data. By construction, the series we thus recover is visually indistinguishable from the series of CLMX. Then, we recomputed the statistics and trend tests with the reverse-engineered data.

Because a portion of the differences in the numerical results from the reverse-engineered data remains,¹³ there is likely a difference in which (some of) the statistics are computed by us and by CLMX. However, because the Vogelsang PS-statistic is straightforward to implement, we conjecture that the remaining small differences are due to some numerical issues. Therefore, we did an additional exercise and simulated a time series with a trend and volatility similar to the observed volatility time-series. Then, we added some small noise to the observations such that, visually, the time series almost look identical. Applying the Vogelsang PS-statistic, we find that we can observe similar differences in the trend estimates like the ones we observe between our estimate and the estimate in CLMX. Hence, although we cannot perfectly replicate the results in CLMX, we believe that we are sufficiently close.

Before we proceed to a more detailed empirical analysis, we discuss two useful properties of the CLMX IV measure. First, we can represent the IV as a difference in average firm and industry variances, i.e.,

$$\begin{aligned}\sigma_{\eta t}^2 &= \sum_i w_{it} \sum_{j \in i} w_{jit} \text{Var}(R_{jit} - R_{it}) \\ &= \sum_i w_{it} \sum_{j \in i} w_{jit} \text{Var}(R_{jit}) - \sum_i w_{it} \text{Var}(R_{it}).\end{aligned}\tag{8}$$

This result requires only that we can classify each element j into its parent category i , and that we use an identical set of weights for portfolio returns and for weighting individual variances. Therefore, the same applies to $\sigma_{\epsilon t}^2$, i.e., the industry component variance in Equation (4), which equals average industry variance minus the market variance. As a consequence, if we leave out the industry layer and measure the IV directly relative to the market, the corresponding IV would equal the sum of firm (idiosyncratic) and industry components in the market–industry–firm setting of CLMX.

¹³Although we do not get precisely the same numbers as in CMLX, we obtain non-significant trends in the market and industry components, and a significant trend in the IV.

To provide additional insights into the determinants of the IV, we consider the following simplified example. We assume that at time t , the industries have an equal number of firms and capital distribution, i.e., $w_{jit} \equiv \bar{w}_{jt}$ for all i . We denote the capitalization weights of individual industries by w_{it} . Further, we assume that for all t , all stocks in industry i have equal variance $\text{Var}(R_{jit}) = \bar{\sigma}_{it}^2$, and that all pairwise correlations within industry are equal to ρ_t . Then, as we show in Appendix B, it holds that

$$\sigma_{\eta t}^2 = \underbrace{\left(\sum_i w_{it} \bar{\sigma}_{it}^2 \right)}_{\text{Average variance}} \underbrace{(1 - \rho_t)}_{\text{Average correlation}} \underbrace{\left(1 - \sum_j \bar{w}_{jt}^2 \right)}_{\text{Industry concentration}}. \quad (9)$$

In this simplified setting, we can separate three primary drivers that influence the IV estimate multiplicatively. The first term is equal to the average firm variance. The second term equals one minus the within-industry correlation ρ . The third term, $(1 - \sum_j \bar{w}_{jt}^2)$, is negatively related to average industry concentration.

The decomposition in Equation (9) provides a very accurate approximation of the exact IV computed by Equation (8). In our sample (1962 to 2016) of 660 monthly observations, the correlation between the exact IV series and its approximation based on realized moments is 0.9968 and 0.9916 in the VW and EW case.¹⁴ The trend estimate of the approximate VW series over the CLMX sample is 6.8% smaller than that of the exact series, but it remains strongly significant. In the EW case, the approximate series's trend estimate is higher by 3.5% compared to the exact EW IV. Given these results, we use Equation (9) as guidance for our line of arguments. We study the contribution of each of the three channels to the CLMX trend in IV and explore whether these contributions are robust when adjusting for microstructure biases.

¹⁴The high correlation values do not stem purely from short-term fluctuations of the IV. If we compute the correlation of moving averages of both series over a five-year window, the correlation is even higher.

3 The three channels of IV measurement

First, we focus on average variances and correlation, whose measurement will potentially be influenced by the microstructure effects laid out in the introduction. Then, we analyze market concentration and its impact on IV.

3.1 Average variances

In Figure 3, we display the time series of the first driver of IV, average variance, both value weighted and equally weighted (analyzed by Xu and Malkiel (2003)). The value-weighted variances (Panel A) based on different frequencies are relatively close to each other, with the most significant gap in the late 1980s and 1990s, i.e., the end of the CLMX sample. The series based on monthly frequency highlights the initial period of abnormally low variance. For the equal-weighted average (Panel B), the gap between the estimates on different frequencies is substantial. The monthly-return series has a visible jump on the inclusion of NASDAQ. However, even after NASDAQ's inclusion, there seems to be a trend in the equal-weighted measure, though it is much weaker than the daily-return realized variances suggest.

As before, we formally test the hypothesis of having no deterministic trend in average variances for the CLMX sample using Vogelsang (1998). However, we depart from CLMX and run the trend tests on the logarithm of the average variances, rather than on their levels. We do so because Vogelsang computes the critical values in the setting with Gaussian innovations. Hence, they are more accurate for log-series whose increments are closer to Gaussian. Table 2 presents the results. In the full CLMX sample (Table 2, Panel A), the point estimates of the trend based on daily-return realized variance (RV) are above the estimates based on monthly returns (RV_m). However, the evidence against the null of no trend is weaker, due to the increased standard errors for daily-return RV stemming from the persistence of the series. On the shorter sample that starts with the inclusion of NASDAQ stocks, no evidence of a trend in value-weighted variance for the monthly frequency emerges. The equal-weighted series based on monthly returns has a significant trend, though its magnitude is less than half of the one found on the daily frequency.

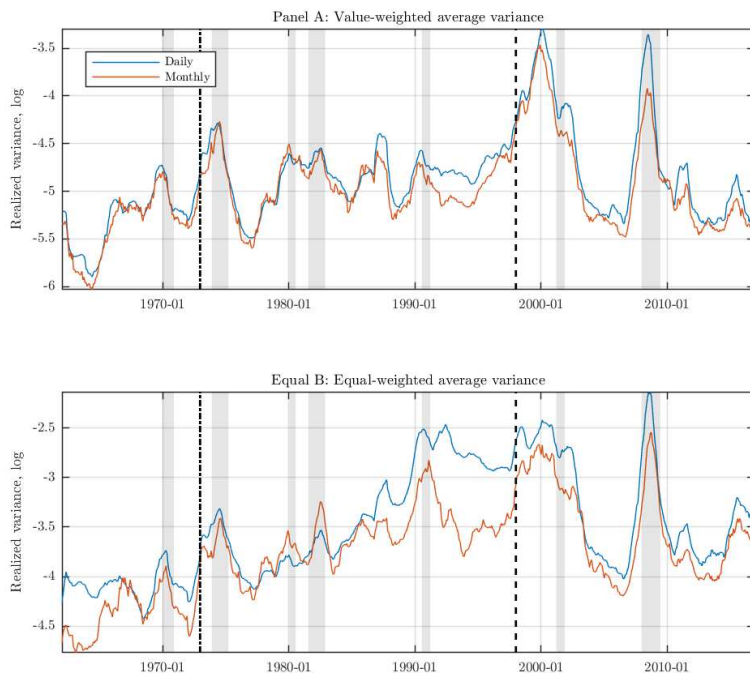


Figure 3: EW and VW variances.

Description: Comparison of equal-weighted and value-weighted variances in log scale. The series correspond to a 12-month moving average. Panel A: Value-weighted average of realized variances. Panel B: Equal-weighted average of realized variances. Shaded areas correspond to NBER recessions. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: Average realized variances of stocks are higher at higher frequencies, more so when weighted equally.

	Full		NASDAQ		Low Price	
	VW	EW	VW	EW	VW	EW
Panel A: 1962-1997, (N = 426)						
RV	1.89 ^{**}	3.47 ^{**}			1.28 [*]	2.49 [*]
RV_m	1.72 [*]	3.31 ^{***}			1.77 ^{***}	2.97 ^{***}
Hasbrouck	1.82 ^{**}	3.29 ^{***}			1.49 ^{***}	2.63 ^{**}
Roll	1.66 [*]	2.70 ^{***}			1.23 ^{**}	2.06 ^{***}
Smoothed	1.71 [*]	2.13 ^{***}			1.28 ^{**}	1.70 ^{***}
Panel B: 1972-1997, (N = 300)						
RV	0.47	4.88 [*]	3.17 ^{**}	6.09 ^{**}	2.82 ^{**}	5.62 ^{**}
RV_m	-0.16	2.34 ^{***}	0.80	2.24 ^{***}	1.28 ^{**}	3.02 ^{***}
Hasbrouck	0.01	3.15 [*]	1.39 [*]	3.38 ^{**}	1.81 ^{**}	3.87 ^{**}
Roll	-0.43	2.32 ^{**}	1.01	2.64 ^{***}	1.15 [*]	3.05 ^{***}
Smoothed	0.04	1.24 [*]	1.08	1.04 [*]	0.97	1.78 ^{**}

^{*}, ^{**}, and ^{***} denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2: Trend tests of average variance series.

Description: Vogelsang $PS^1 - t$ test of average variance series. All values are multiplied by 1,000. We indicate whether the individual variances are weighted equally (EW) or by their capitalization (VW). The Full series contains all stocks that have a minimum of ten daily observations in given month. NASDAQ columns consider only a subset of stocks listed on NASDAQ. Low Price columns are based on stocks with price no greater than 10\$. RV and RV_m are the realized variances based on one month of daily and monthly returns, where the latter uses a single observation. Hasbrouck and Roll series correspond to the MCMC (Hasbrouck, 2009) and moment estimators of the model of Roll, 1984. Smoothed series is the outcome of the filtering procedure for the model described in Section 3.1.2.

Interpretation: There is a significant trend in the average variance over 1962-1997 period. The evidence of a trend is much weaker when we account for bid-ask bounce and price discreteness.

3.1.1 Adjusting for the bid-ask bounce using Roll (1984)

We can link the difference between realized variances in Figure 3 to the sample autocorrelation in daily stock returns. Indeed, we find that the average daily autocorrelation becomes strongly negative from the mid-1980s till the end of the CLMX sample. This observation is consistent with the bid-ask bounce model of Roll (1984). In this model, the efficient log-price process (s_t^*) is assumed to follow a random walk, and the observed log-price (s) corresponds to the bid or the ask with equal probability. Formally,

$$s_t = \begin{cases} s_t^* + c, & \text{prob} = 0.5, \\ s_t^* - c, & \text{prob} = 0.5, \end{cases} \quad \text{with} \quad s_t^* = s_{t-1}^* + \sigma \epsilon_t. \quad (10)$$

Under these assumptions, the sample variance is biased upward, and the first-order return autocovariance equals $-c^2$ (higher orders equal zero). To correct for the bias in estimates of the sample variance, we use the model-implied relation

$$\text{Var}(r_t) + 2 \text{Cov}(r_t, r_{t-1}) = \sigma^2. \quad (11)$$

A common approach to estimate the underlying variance σ^2 is to plug sample moments into the left-hand side of Equation (11). One caveat is that due to finite-sample variation, the resulting variance estimate might be negative, even if the model is correctly specified. As a crude way to ensure the non-negativeness of the variance, we define

$$\hat{\sigma}_{\text{Roll}}^2 = \max(\widehat{\text{Var}}(r_t) + 2\widehat{\text{Cov}}(r_t, r_{t-1}), 0). \quad (12)$$

We also estimate Roll's model by means of the Gibbs sampler as proposed by Hasbrouck (2009), which ensures positive estimates of variance. We use both methods to obtain monthly estimates of return volatilities for all stocks in the universe.¹⁵

¹⁵Corwin and Schultz (2012) develop another estimator of the volatility and the spread which uses daily high and low prices, instead of autocorrelations. However, we do not consider this approach here, because the CRSP database does not contain high and low prices for all stocks in all periods. Moreover, as argued in the Online Appendix, the estimator of Corwin and Schultz (2012) may introduce a nontrivial bias.

3.1.2 Adjusting for additional microstructure biases

Even though Roll's model generates bid-ask bounce, it implies that observed quotes vary smoothly with the efficient price, so it does not generate flat price segments. In addition, there are other microstructure and data-related issues with daily CRSP returns. First, price discreteness may affect the estimates of the variance, as illustrated in Figure 1. Indeed, Bali *et al.* (2008) find that the trend in the IV, which was detected over a period of minimum tick size of USD $1/8$, was concentrated among NASDAQ stocks with low price. Because these stocks are the ones most affected by price discreteness, it is critical to understand its role in the measurement of volatility. A second concern is that both volatility and bid-ask spreads vary over time. Bid-ask spread often oscillates on a day-to-day basis. Thus, holding it fixed over the given month is inadequate. Our third concern is related to the fact that prices in the CRSP database have been recorded in two different ways: as the quote midpoint or as the closing price. Furthermore, the quotes can correspond to either closing quotes or inside quotes, depending on the exchange and period under consideration.¹⁶ For a consistent comparison of stock variances across time and exchanges, it is therefore essential to accommodate these differences.

To take into account the aforementioned microstructure effects and features of the data, we introduce the following price model. We denote the true, efficient log-price by s^* and use capital letters for level variables. Then, we assume that the efficient (log-) bid b and ask a are symmetric around the efficient log-price, that is,

$$a_t^* = s_t^* + C_t, \quad b_t^* = s_t^* - C_t, \quad (13)$$

where $C_t > 0$ denotes (one-half of) the time-varying bid-ask spread. Next, we assume that the observed quotes are a rounded version of their efficient counterparts, with bid rounded down and ask rounded up. If available, we use quotes instead of the recorded closing price. If they are unavailable, as is the case for NYSE stocks before 1992, we use the closing price for the estimation, and assume that it equals the bid and ask with equal probability.

¹⁶Section A of the Online Appendix summarizes these features in more detail and discusses their potential effect on the measurement of variance.

Formally,

$$S_t = \begin{cases} A_t = \left\lceil \frac{A^*}{d_t} \right\rceil d_t, & \text{prob} = 0.5, \\ B_t = \left\lfloor \frac{B^*}{d_t} \right\rfloor d_t, & \text{prob} = 0.5, \end{cases} \quad (14)$$

where d_t is the effective tick size (e.g., USD $1/8$) prevailing in period t . To complete the model, we specify the transition probabilities as

$$\log(S_{t+1}^* F_{t+1} + X_{t+1}) = s_t^* + \sigma_{t+1} \epsilon_{s,t+1}, \quad (15)$$

$$\log(\sigma_{t+1}) = \log(\bar{\sigma}) + \varphi_\sigma(\log(\sigma_t) - \log(\bar{\sigma})) + \gamma_\sigma \epsilon_{\sigma,t+1}, \quad (16)$$

$$c_{t+1} = \chi_{t+1} + \gamma_c \epsilon_{c,t+1}, \quad (17)$$

$$\chi_{t+1} = \bar{\chi} + \varphi_\chi(\chi_t - \bar{\chi}) + \gamma_\chi \epsilon_{\chi,t+1}, \quad (18)$$

with iid Gaussian innovation terms $\epsilon_{\cdot,t}$, independent of each other. By X_t , we denote the dividend in period t , and by F_t , we mean the adjustment factor reflecting events such as stock splits. The structure of the spread is motivated by the typical evolution of a spread, characterized by a slowly varying mean and a noise component. We estimate the states through particle filtering.¹⁷ Although we impose independence of innovations in the transition densities of spread and volatility and estimate each stocks' states separately, the posterior means of those variables can be correlated, both over time and across assets. Consistent with the model of Bollen *et al.* (2004), the smoothed volatility estimates, which are clear of bid-ask bounce effects, are positively correlated with the spread in the cross-section of stocks. Moreover, the spread is positively related to the difference between the realized variance and the smoothed estimates, which we can interpret as a measure of bias in the realized variance. However, this relation holds only on subsets where trading prices are available.

3.1.3 Is the trend in average variance robust?

Figure 4 compares the estimates of average variance using different estimation methods.¹⁸ For the equal-weighted average, all the robust measures of average variance lie below the sample estimate based on daily

¹⁷For the complete data treatment and details of the filtering procedure, we refer to Appendix C.

¹⁸The measures of Roll, Hasbrouck, and the smoothed estimates all correspond to the variance of monthly log-returns, while the CLMX measure is based on simple returns.

returns, with the largest difference in the period from 1985 to 2000. These differences largely disappeared with the quote decimalization in 2001. Both estimates of Roll's model are rather close to squared monthly returns. The smoothed estimates are even lower. One possible explanation of the difference is that the smoothed series takes into account price discreteness and spread volatility, which are not entirely eliminated when using a monthly frequency. Another possibility is that the difference stems from the fact that our filtering procedure eliminates the outliers. Yet another possible explanation is that part of the difference may stem from our use of quotes rather than closing prices, combined with the fact that some closing prices fall outside of the bid-ask range. In Figure C1 of the appendix, we show that most of the observations flagged as outliers occur after 1995, so the elimination of outliers is unlikely to be the main driving force of the gap. The explanation based on using quotes instead of closing prices seems unlikely due to the following observation. Figure 4 shows that the gap is present even in 1972-1982 period, when NYSE stocks' quotes and NASDAQ stocks' closing prices are unavailable, so the data used by different methods are identical. Combined with the fact that the gap closes in period 1997-2001 when the exchanges undergone multiple tick-size reductions, indicates that price discreteness is the most likely source. We leave an in-depth investigation of its source for future research. For the value-weighted IV, the differences are quantitatively smaller, because it puts lower weight on small stocks, which are more heavily affected by microstructure effects.

The results of Vogelsang's trend test, reported in the left column of Table 2, are broadly consistent with the observed patterns: point estimates for the robust measures are in the majority of cases below the realized variance estimated from daily data. Still, they result in a rejection of the no-trend hypothesis in most cases. In Figure 3, we see that the volatility

The correspondence between the variance of a random variable and its (nonlinear) transform is, in general, not trivial. With conditionally Gaussian log-returns with zero mean, the corresponding variance of simple returns (i.e., log-normal random variables) can be computed as $\text{Var}(R) = e^{\sigma^2}(e^{\sigma^2} - 1)$, where σ^2 denotes the variance of log-returns. We used the Gaussian case as an approximation for the variance of simple returns. Using the simple return variance approximation does not alter our results substantially, as long as we bound the monthly variance of log-returns by a (large) constant, for example, by unity. Without such a bound, some extreme variances of log-returns blow up the estimates due to exponentiation. These extreme values often represent a data error (see, Online Appendix, Section B), or occur for penny stocks.

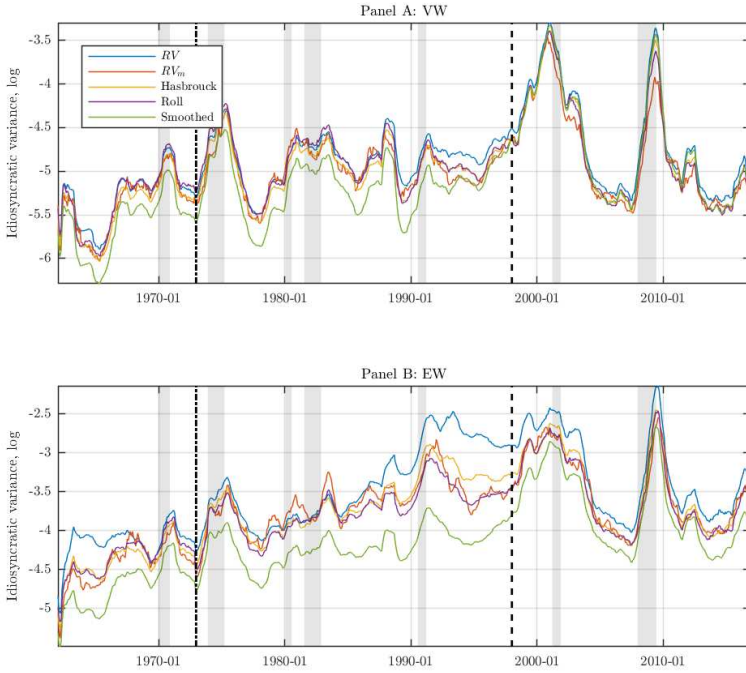


Figure 4: Average variance: Comparison of estimation methods.

Description: Equally weighted (Panel A) and value-weighted (Panel B) average variance, while RV and RV_m are the realized variance measures of daily returns and the squared monthly return. The Roll series is the sample moment estimate of the model of Roll (1984). “Hasbrouck” denotes the Gibbs-sampler estimate of Roll’s model proposed by Hasbrouck (2009). The smoothed series is obtained from the model described in Section 3.1.2. The values are shown in log scale, smoothed by a moving average filter with a window of 12 months. Shaded areas correspond to NBER recessions. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: When we account for microstructure biases, the average variance is lower and with a weaker or even no trend.

is abnormally low at the very beginning of the sample period, which can drive the trend tests. To verify this observation, we repeat the battery of trend tests on the CLMX sample, truncated to December 1972 (inclusion of NASDAQ) from the beginning. In this restricted period (Panel B), we see that the equal-weighted daily-return RV has an even stronger trend than in the full CLMX sample, and the gap between the baseline and robust measures becomes immense. The confidence interval for the trend estimate for the smoothed variance series contains zero.

Because Bali *et al.* (2008) find that the trend observed in the CLMX period is concentrated among low-price stocks and NASDAQ, we investigate to what extent the trend in average variance is robust for those subgroups. The results in Table 2 suggest that, while the sample variance shows strong evidence of a trend, the evidence becomes weaker when we adjust for microstructure biases. For the smoothed estimates, the absence of a trend cannot be rejected, except for the equal-weighted measure for stocks with a low price. Hence, our result that the trend in average variances disappears remains valid on the NASDAQ subsample.

3.2 Correlation

For our second channel driving IV, average correlation, we observe from Equation (8) that IV is negatively related to industry variances, which are positively associated with individual stocks' variances and within-industry correlations. Hence, average correlation is negatively related to IV. We start our empirical analysis of correlations by estimating their average across all pairs of stocks in our sample. We compute both equal- and value-weighted averages. Every month, we calculate the correlation based on one year of data and use only stocks with a full year of observations, eliminating any stocks listed or delisted during the year, or that have any missing observations. We estimate the correlations using daily, weekly, and monthly returns.

Panels A and C of Figure 5 confirm the observation of CLMX for their time period. The correlations based on a higher (daily) frequency are much lower than those obtained from monthly data. This observation holds for the entire CLMX sample, both for the value-weighted (Panel A) and equal-weighted (Panel C) cases. Because we use an identical set of firms

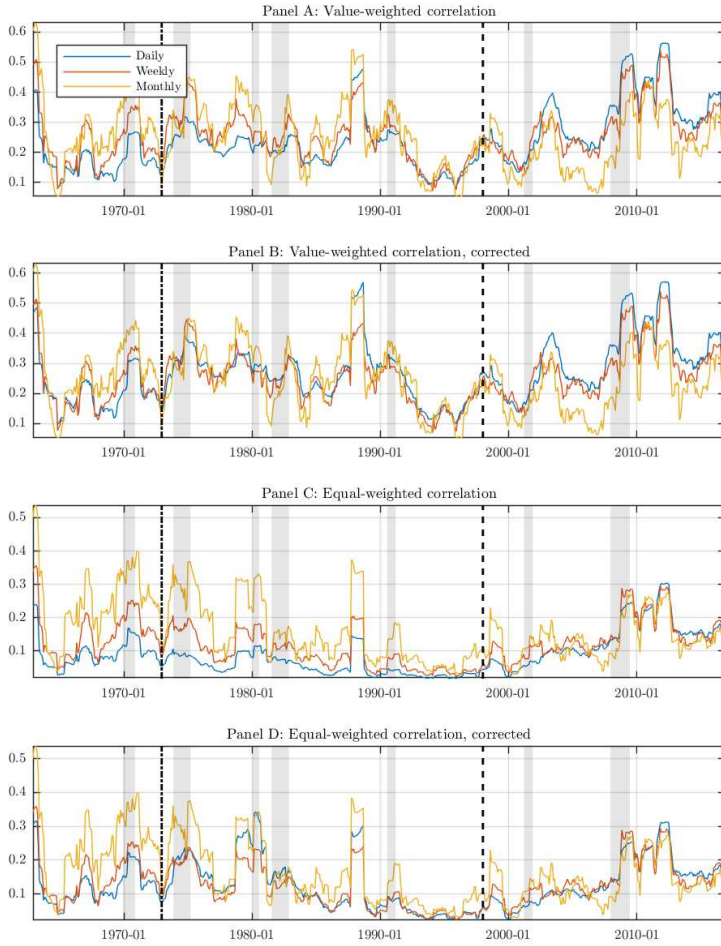


Figure 5: Average correlations.

Description: Correlations between all pairs of stocks, computed using one year of data at daily, weekly and monthly frequencies. In Panel A, we plot the value-weighted average correlation. In Panel B, we plot the corrected value-weighted correlations with $w_{ij} \propto V_i V_j$. In Panels C and D, we plot the corresponding equal-weighted correlations. Shaded areas correspond to NBER recessions. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: Average sample correlation is higher at lower frequencies in the CLMX sample. Asynchronicity correction based on frequency of non-trades (zero returns) provides a reasonable approximation of the bias.

for all frequencies, the only reason for these differences must stem from the negative (cross-) autocorrelations of daily returns in the CLMX sample. Interestingly, the wedge between the correlations calculated at different frequencies disappears in the post-CLMX period.

These observations seem difficult to justify on the grounds of changing fundamental factors. We argue that at least three microstructure effects may cause severe correlation biases. First, an asynchronicity bias arises because the computation of the sample covariance requires pairs of returns covering the same period, while in practice, trades of different assets occur at different times. A simple treatment is to “synchronize” the prices to a fixed grid by using the last known price value. This procedure biases the sample correlation towards zero.¹⁹ In the extreme case, when no trade occurs between two adjacent sampling points, the price is flat (zero return), and the resulting contribution to the sample covariance is zero. Second, price discreteness has a similar effect and, when the tick size is large compared to the price of the stock, produces flat prices. Third, microstructure noise due to the bid-ask bounce inflates the variance of the individual stocks but keeps the covariances intact as long as the noise is independent across assets. In turn, the correlations are again biased towards zero.

For simplicity, we label all zero-return observations as “non-trades,” even though zero returns may also occur for nonzero trading volumes. For stocks with zero trading volume, as well as for those with unavailable trading prices, CRSP records a closing-quote midpoint. The latter applies to all NASDAQ stocks before November 1, 1982, and for NASDAQ SmallCap Market before June 15, 1992. Quote midpoints remain flat unless the bid or ask (inside or closing) quote is revised or filled. Filling is less likely for illiquid stocks with wide bid-ask spreads.

Price discreteness further widens the spread,²⁰ causing more frequent non-trades. It also generates non-trades for closing prices, because the round-off to the nearest tick makes the probability of a zero return positive, even if the underlying distribution of returns has continuous support. For NASDAQ stocks, Christie and Schultz (1994) show that when the tick size

¹⁹The phenomenon of decreasing return correlations with increasing frequency is commonly referred to as the Epps effect. See Epps (1979).

²⁰See Hasbrouck (1999), Zhang *et al.* (2008), and Bessembinder (2000).

was USD $1/8$, NASDAQ dealers avoided quotes ending with odd eighths, amplifying price discreteness.²¹ The bid-ask bounce has the opposite effect on the frequency of non-trades because the trading prices would move even if the efficient price were constant.

To gauge the potential magnitude of the correlation bias, we compute the percentage of non-trades over time. The equal-weighted percentage of non-trades peaked around 50% right after the inclusion of NASDAQ and started to decrease with tick-size reductions, beginning in 1992 with Amex and ending with the decimalization on all exchanges in 2001.²² Comparing with the value-weighted share, we find that non-trades are concentrated among smaller stocks. For the entire CMLX period, the average equal-weighted and value-weighted monthly shares of zero trades are around 32% and 13%, respectively. After the introduction of decimalization in 2001, these numbers drop to 4% and 1%. Because each non-trade affects all sample correlations with a given stock, the share of non-trades may have a significant impact on portfolio variances, as all return cross-products vanish.

To obtain a better estimate of the asynchronicity bias, we consider the following simplified setup. We assume that non-trades for asset i occur with probability λ_i , independently of the time and other assets' non-trades and returns. Then, we define the bias correction term B_{ij} as

$$\hat{\rho}_{ij} := \rho_{ij} B_{ij}(\lambda_i, \lambda_j) \approx \rho_{ij} B_i(\lambda_i) B_j(\lambda_j). \quad (19)$$

Hence, we can interpret $1 - B_{ij}$ as the percentage bias in the sample correlation. To facilitate efficient computation of the bias in large portfolios, it is useful to approximate the bias correction by $B_{ij}(\lambda_i, \lambda_j) \approx B_i(\lambda_i) B_j(\lambda_j)$, separating the effect of individual non-trades. A naive estimate of B_{ij} corresponds to

$$B_{ij}^N \approx (1 - \lambda_i)(1 - \lambda_j), \quad (20)$$

which is the expected percentage of return cross-products affected by a non-trade. The naive correction works adequately for moderate shares of

²¹In the Online Appendix, Section A, we show that odd-eighth avoidance is valid over the entire period before quote "decimalization".

²²The equal-weighted series is computed as the proportion cross-firm average percentage of non-trades within a month. For the value-weighted series, we attach a weight to each firm's non-trade share according to its market capitalization.

non-trades but deteriorates for large ones, for which B^N is biased upward. A better approximation can be obtained by simulation. We sample a bi-variate random walk for a price process and generate iid non-trades with different probabilities λ . Then, we estimate the sample correlations based on observations contaminated by non-trades. Finally, we regress $\log(\hat{\rho}/\rho)$ on $\log(1 - \lambda_i) + \log(1 - \lambda_j)$, which yields an estimate of 0.8. Thus, we use $B_i = (1 - \lambda_i)^{4/5}$.

In what follows, we consider the variances and non-trade probabilities λ to be known.²³ In our implementation, we replace them with their sample estimates. In the first step, we compute the average bias in the pairwise correlations $1 - \bar{B}_w$ (with weights w), where we define \bar{B}_w as

$$\bar{B}_w = \frac{\sum_{i \neq j}^N w_i w_j B_{ij}}{\sum_{i \neq j}^N w_i w_j}. \quad (21)$$

The average percentage bias in the correlations $(1 - \bar{B})$, using the naive bias estimates, goes up to 75% in the equal-weighted case and 30% in the value-weighted case at the time when the non-trade shares are peaking. Because the variance of a diversified portfolio is dominated by the average covariance of its constituents, we can view the average bias in the sample correlation as a crude estimate of the percentage bias in the sample variance. However, such an estimate neglects the cross-sectional correlation between the bias B_{ij} and the weight in the portfolio variance $w_i w_j \hat{\sigma}_{ij}$. For a more accurate approximation, we apply the bias correction in Equation (19) directly to the individual correlations, and compute their average via

$$\bar{\rho}_C(w) = \frac{\sum_{i \neq j} w_i w_j \rho_{ij}}{\sum_{i \neq j} w_i w_j} \approx \frac{\sum_{i \neq j} \frac{w_i w_j \hat{\rho}_{ij}}{\bar{B}_i \bar{B}_j}}{\sum_{i \neq j} w_i w_j} = \frac{\widehat{\text{Var}}\left(\sum_i \frac{w_i R_i}{\bar{\sigma}_i \bar{B}_i}\right) - \sum_i \frac{w_i^2}{\bar{B}_i^2}}{1 - \sum_i w_i^2}, \quad (22)$$

where we use “ \wedge ” to denote the sample moments.²⁴ Analogously, we

²³An additive microstructure noise, which is independent of the returns and the asset, would not affect estimates of the covariance, but will still bias upward the variances in the denominator of the correlation. Therefore, the correlation estimates will remain biased toward zero, and our bias correction can be deemed conservative.

²⁴We exclude stocks with zero sample variance, for which the sample correlation is ill-defined.

compute the bias-corrected portfolio variances by²⁵

$$\text{Var}_C \left(\sum_i w_i R_i \right) \approx \sum_i w_i^2 \hat{\sigma}_i^2 + \widehat{\text{Var}} \left(\sum_i \frac{w_i R_i}{B_i} \right) - \sum_i \frac{w_i^2 \hat{\sigma}_i^2}{B_i^2}. \quad (23)$$

In Figure 5, Panels C and D, we present the average correlations corrected for the asynchronicity bias as defined in Equation (22). The bias-corrected average correlations in the recent period are rather close to the levels observed in the 1970s and early 1980s. This observation holds even for the equal-weighted average, which is not the case for the uncorrected estimates. The only period of lower correlations surrounds the dot-com bubble. Also, the bias-corrected estimates are similar to the estimates based on monthly returns, where the effects of the microstructure are less significant, which generates confidence in the bias correction procedure. Part of the difference between the bias-corrected estimates based on different frequencies can be caused by effects of the microstructure not reflected in non-trades, for example, the bid-ask bounce and price discreteness. The transitory drop in correlations based on monthly data could also be partially driven by residual microstructure effects for highly illiquid stocks. In the full sample, where we do not exclude stocks without a full year of data, the bias correction is probably stronger, because the excluded stocks have a higher non-trade frequency on average.

As the above analysis indicates, the realized correlations based on daily CRSP return data are biased towards zero in the early part of the sample, but the biases vanish with quote decimalization. These biases reconcile the differences between average correlations based on different sampling frequencies, as observed by CLMX in their Figure 5. Once we correct for the asynchronicity bias in correlation, the long-term mean of correlations stabilizes. Therefore, the break in correlation biases associated with decimalization contributes to the reversal of the IV, even more so for the equal-weighted measure.²⁶

²⁵In principle, the corrected estimates can violate positive semidefiniteness of the covariance matrix, leading to an infeasible correlation or negative variance estimates. However, this issue does not occur in our application, so we decide to disregard it.

²⁶Because the realized correlations are at their historical low in the 1990s and are negatively related to the IV, they represent another omitted variable in the regression approach of LPSZ. As a consequence, they overstate the role of the bid-ask bounce at the expense of other biases in correlations.

3.3 Market concentration

Theoretically, Equation (9) connotes a negative impact of market concentration on measured IVs. Intuitively, this impact stems from a stronger diversification effect in industry variances, driving the IV up. We consider two measures of market concentration. The first measure is a market-wide Herfindahl–Hirschman index \mathcal{H} ,

$$\mathcal{H} = \sum_j w_j^2, \quad (24)$$

with $w_j = w_i w_{ji}$ denoting the capitalization-based weight of firm j in the market. The second measure is a weighted industry concentration index $\tilde{\mathcal{H}}$,

$$\tilde{\mathcal{H}} = \sum_i w_i \sum_{j \in i} w_{ji}^2, \quad (25)$$

with w_i denoting the weight of industry i and w_{ji} the weight of firm j in industry i . The former is more appropriate if we want to measure the diversification effect in the market portfolio, which affects market variance and also industry-layer variance. The latter is more suitable for studying an average diversification effect in industry portfolios, whose variance drives the industry-layer variance and the IV.

In Figure 6, we plot the value-weighted and equal-weighted concentration indices. In Panel A, the period of the increase in value-weighted IV as perceived by CLMX coincides with a period of decreasing market and industry concentration. The concentration stabilizes in the post-CLMX period when the IV in Figure 2 tends to decline again. A similar pattern emerges when we assign equal weights to the firms. In Panel B, the decrease in the concentration based on equal weights is negatively related to the number of listed firms in the market and industries. Indeed, in the CLMX period, we witness a substantial increase in the number of listed firms and a decrease afterward.²⁷

In the simplified setting of Equation (9), the expected squared IV depends multiplicatively on the industry concentration via $1 - \tilde{\mathcal{H}}$. For the

²⁷Doidge *et al.* (2017) provide an excellent discussion of the causes for the changing trend in the number of listed firms, highlighting a decreased propensity to be listed, partially explained by regulatory changes and developments in financial markets.

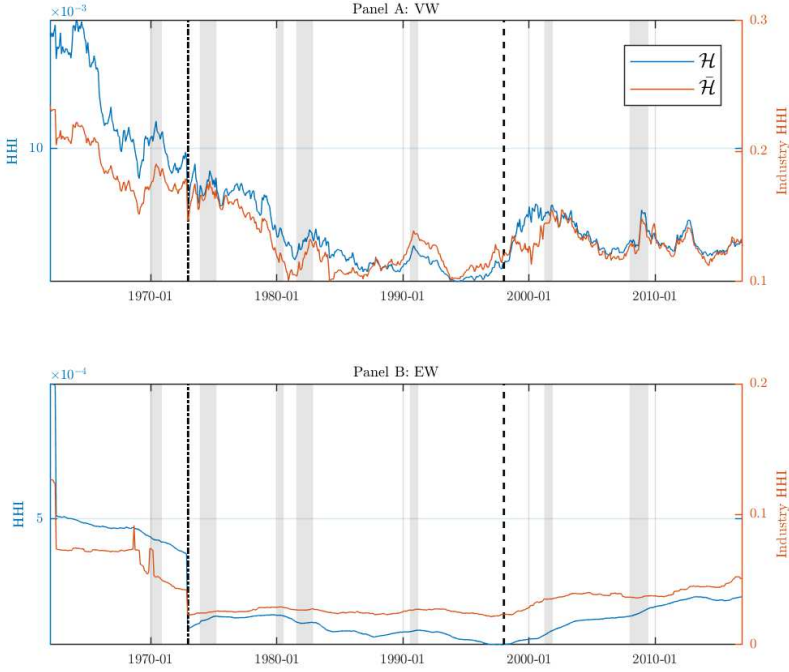


Figure 6: Market concentration time series.

Description: Panel A shows the time series of value-weighted market concentration measured by the Herfindahl–Hirschman index (\mathcal{H}), defined in Equation (24), and a weighted average of industry concentrations ($\bar{\mathcal{H}}$), defined in Equation (25). Panel B shows the market and weighted industry concentration based on equal weights, that is, $w_i \equiv \frac{1}{F}$, where F is the number of firms in a given period. All values are computed at the beginning of the month. The sample spans the period from January 1962 to December 2016. Shaded areas correspond to NBER recessions. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: Equity market concentration decreased over the CLMX sample and thus contributes to the upward trend in the IV.

value-weighted IV in the CLMX period, $1 - \mathcal{H}$ increased by approximately 17.8% peak-to-trough. At the same time, IV based on realized variances roughly doubled. Hence, industry concentration plays only a secondary role in microstructure biases. However, the size of its effect remains economically significant.²⁸ The effect of industry concentration on equal-weighted IV is negligible because equal weights minimize Herfindahl–Hirschman index.

4 Was there a trend in the IV?

Studying the different channels driving the IV in the previous section, we have learned the following. First, the increase in average realized variance from 1962 to 1997 is, to no small extent, an artifact of the bid-ask bounce, further amplified by price discreteness. Second, these effects, combined with price asynchronicity, bias the sample (realized) correlations towards zero. Third, after the quote decimalization in 2001, these microstructure effects in daily returns become negligible. Hence, average variances and correlations calculated with different sampling frequencies converge. Lastly, the decrease in industry concentration over the period covered by CLMX explains, at most, only a part of the trend in the IV.

To finalize our analysis of a potential trend in IV during the CLMX period, we now correct for the biases in the estimates of the variance and correlation.²⁹ Before presenting the test statistics, Figure 7 visualizes the differences of the bias-corrected IV and baseline IV series in Equation (9). The value-weighted baseline IV is trending in the CLMX sample and reverts afterward. Both patterns are (visually) weaker for the Gibbs estimate of

²⁸Clearly, without controlling for industry correlation, the explanation of the IV trend is incomplete. Hence, as we argue in the Online Appendix, Section B, the argument of LPSZ that the bid-ask bounce alone explains the trend in value-weighted IV does not appear to be pertinent due to an omitted variable bias in their regression tests, hence overstating the role of the bid-ask bounce.

²⁹To obtain an indication of the impact of microstructure effects on the measurement of the IV, we also analyzed IV estimates based on different sampling frequencies. The results were in line with the microstructure bias rationale, i.e., the IV based on a lower frequency, which is (more) robust to asynchronicity and bid-ask bounce effects, falls below the series obtained using higher frequencies. This pattern holds particularly in periods with strong microstructure effects. These results can be obtained from the authors.

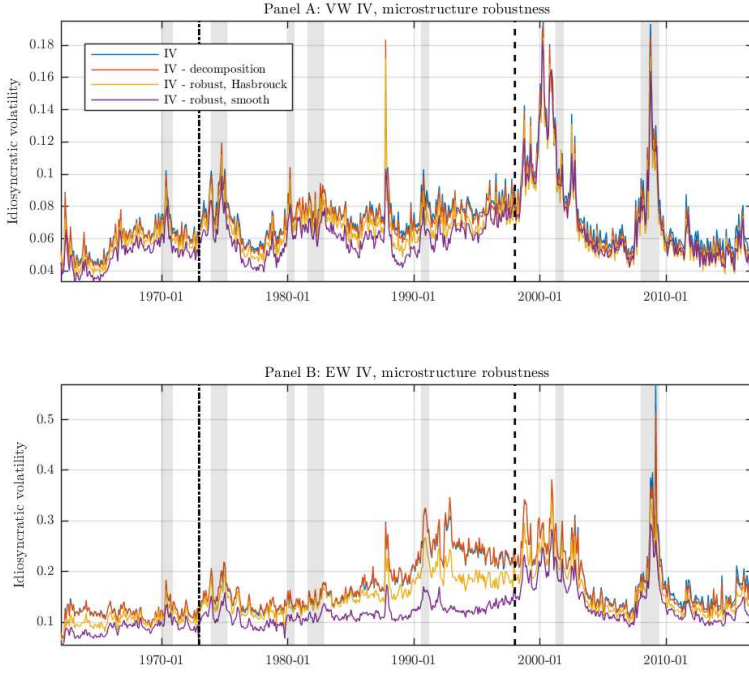


Figure 7: Original and bias-corrected estimates of the IV.

Description: Comparison of the original estimates of the IV, the robust IV estimates following Hasbrouck (2009), and our IV estimate using the robust variance measures defined in Section 3.1 and adjusting the correlation for asynchronicity bias via Equation (22). In Panel A, we plot the value-weighted IV and in Panel B the equal-weighted IV. Dot-dashed vertical line, corresponding to December 14, 1972, marks the inclusion of NASDAQ stocks into the sample. The dashed vertical line denotes the end of the CLMX sample (December 31, 1997).

Interpretation: Once we account for bid-ask bounce and asynchronicity, the trend in IV is weaker and the reversal is essentially gone.

Roll's model as proposed by Hasbrouck (2009), and virtually non-existent for the smoothed estimates described in Section 3.1. In the equal-weighted case, the ordering of the estimates is similar. Here, the smoothed estimates still show signs of an upward trend before 1997, though its magnitude is much weaker than in the baseline case.

Table 3 shows the tests of the hypothesis of no trend for the full CLMX sample and its subset starting in 1972. In the full sample, the magnitudes of the trend are comparable, due to the low values of the robust variances at the start of the sample. In the restricted sample, the estimates of the magnitude of trend drop by 20% to 40% for the value-weighted measures, and by 25% to 60% for the equal-weighted averages. The estimates for IV based on smoothed variances are visibly increasing towards the end of the CLMX period, which falls in the buildup of the dot-com bubble. We have shown that, when considering each exchange separately, the robust measures of average variance show little evidence of a trend, but the average variance of NASDAQ is shifted upward. Furthermore, during the period covered by CLMX, NASDAQ's proportion of all exchange-listed stocks in the sample increased markedly, while its share of capitalization grew at a rather slow pace. As a consequence, the growth in the equal-weighted average variance corresponds to the increasing prominence of NASDAQ stocks in our universe of stocks.

When we restrict our attention to the period after the inclusion of NASDAQ, namely, from 1972 to 1997, we find that the trend for the value-weighted IV becomes statistically insignificant, also when using the method of CLMX. Hence, the strength of evidence of a trend in IV is sensitive to the inclusion of first few years in the sample. Lastly, in the right columns of Table 3 we show the test results for series net of industry concentration effects, computed as $IV/(1 - \bar{\mathcal{H}})$. Once the industry concentration effects are turned off, the value-weighted trend estimates are even weaker. The rightmost column shows the trend coefficients for the IV share, computed as

$$\text{IV share} = \frac{\sigma_{\epsilon t}^2}{\sigma_{\epsilon t}^2 + \sigma_{\eta t}^2 + \sigma_{mt}^2} = \frac{\text{Idiosyncratic variance}}{\text{Average Variance}}. \quad (26)$$

The trend tests for the IV share show that the absence of trend in it cannot be rejected in the baseline case of the CLMX approach. Therefore, the interpretation of the CLMX results as an evidence of an increasing promi-

nence of the (VW) idiosyncratic component relative to systemic ones is not justified. The hypothesis is rejected for the EW IV share, due to the decrease in correlations at the end of the CLMX sample when the number of listed companies peaked.

Overall, the evidence of a trend in the IV series and its reversal is limited. The primary driver of the trend is the bias in variance due to bid-ask bounce effects, which materialized with the increasing availability of closing prices for NASDAQ stocks in the CRSP database in the 1980s and 1990s, and became negligible after quote decimalization, which reduced the size of effective spreads and the bias itself. In contrast to the conclusions of LPSZ for the value-weighted IV, we stress that the bid-ask bounce is not the only driving force. The decimalization also eliminated biases in correlation estimates due to price discreteness and asynchronicity. These biased correlation estimates, in turn, bias the IV estimates upwards. Third, changes in industry concentration account for approximately 18% of the trend in idiosyncratic variance in the value-weighted case. We also show that equal-weighted IV retains a weak trend due to the increasing proportion of the more volatile NASDAQ stocks.

5 Conclusion

In this study, we revisit the question of the existence of an increase in the IV in the period from 1962 to 2000, perceived as a trend in earlier studies, and discuss its primary drivers. We separately study three channels driving the IV: average variance, average within-industry correlation, and industry concentration. We show that before the quote decimalization in 2001, the estimates of the realized variance based on daily returns are severely biased upward. We identify the bid-ask bounce and price discreteness as the leading causes. These microstructure effects are the strongest for NASDAQ stocks with a low price, which Bali *et al.* (2008) show drove the increase in the IV. Because in the 1980s and 1990s, these stocks grew in numbers, and also their trading prices (prone to bid-ask bounce) became increasingly available, the severity of the biases increased over time, until they were wiped out by tick-size reductions.

To deal with microstructure biases, we design a model that deals explicitly with the price discreteness, bid-ask bounce, and other features

	Original		Net of IC		IV Share	
	VW	EW	VW	EW	VW	EW
Panel A: 1962-1997, (N = 426)						
Original	1.94 ^{**}	3.62 ^{**}	1.59 ^{**}	3.46 [*]	0.03	0.20 ^{**}
Hasbrouck	1.73 ^{**}	3.54 ^{***}	1.38 [*]	3.38 ^{***}	-0.06	0.22 [*]
Smoothed	1.61 [*]	2.38 ^{***}	1.27 [*]	2.22 ^{***}	-0.06	0.22 [*]
Panel B: 1972-1997, (N = 300)						
Original	0.90	5.08 [*]	0.66	5.08 [*]	0.29 ^{**}	0.18 ^{***}
Hasbrouck	0.46	3.79 [*]	0.22	3.79 [*]	0.29 [*]	0.53 ^{***}
Smoothed	0.50	1.88 [*]	0.26	1.88 [*]	0.29 [*]	0.53 ^{***}

*, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3: Trend tests of microstructure-adjusted idiosyncratic volatility.

Description: Vogelsang $PS^1 - t$ test of idiosyncratic volatility series corrected for microstructure biases. All values are multiplied by 1,000. We indicate whether the individual variances are weighted equally (EW) or by their capitalization (VW). Original refers to the estimates of CLMX based on sample moments and daily returns. Hasbrouck and Smoothed series use the approximation of the IV series in Equation (9), and the bias-corrected average correlations defined in (22). Hasbrouck series then applies the MCMC estimates of the Roll's model proposed by Hasbrouck (2009). Smoothed series is the outcome of the filtering procedure for the model described in Section 3.1. Panel labeled Original presents results for the log-IV, in Net of IC we present results for the IV scaled by $(1 - \mathcal{H})$, i.e., net of industry concentration effects as implied by Equation (9). IV Share denotes the ratio of Idiosyncratic Variance to Total (Average) Variance, where Total Variance equals the sum of idiosyncratic, industry and market components.

Interpretation: Once we correct for microstructure biases, the evidence in favor of a trend in idiosyncratic volatility is weaker and sensitive to inclusion of the first few years in the sample (before NASDAQ).

contained in the CRSP data. When we replace realized variances with variances from our model that are robust to microstructure effects, the trend estimate for average variance decreases by 10% and 39% in the value-weighted and equal-weighted case, respectively. The effect is substantially stronger when we restrict the sample 1972-1997, i.e., since NASDAQ joined the CRSP database till the end of the CLMX sample, with trend estimates lower by 92% in the value-weighted case and by 75% in the equal-weighted case. A similar pattern holds for NASDAQ, and low-price subsamples, which Bali *et al.*, 2008 show to drive the trend in the IV: There is virtually no decrease on the full CLMX sample for the VW average variance. For the EW average, the decrease corresponds to 32% of the CLMX estimate. On the shorter sample of 1972-1997, the trend estimates for both subsamples (NASDAQ, low-price) and weighting schemes (VW, EW) shrink by at least 65%.

Sample correlations based on daily returns are also biased before quote decimalization due to the effects of asynchronicity, price discreteness, and the bid-ask bounce. We confirm the finding of CLMX that equal-weighted correlations estimated using a lower frequency are much higher than those based on daily data. The magnitude of the bias is closely related to the share of zero returns in the sample, which is one of the symptoms of microstructure effects. The decrease in industry concentration from 1962 to 2000 also contributes to a higher (CLMX measure of) IV due to the stronger diversification of industry and market portfolios. Still, its effect is quantitatively small compared to other channels.

Correcting for the bias in variances and correlations, the IV trend estimates for the CLMX sample (1962-1997) decreases by 17% for the value-weighted and by 34% for the equal-weighted IV. On the shorter sample starting in 1972 (with the inclusion of NASDAQ), the microstructure effects constitute 44% (VW case) and 62% (EW case) of the trend estimates. The most striking feature is the abnormally low IV at the beginning of the sample period. Lastly, with our microstructure bias adjustments, the IV series' trend reversal after 1997 is no longer visible.

References

- Aït-Sahalia, Y. and J. Yu (Mar. 2009). “High frequency market microstructure noise estimates and liquidity measures”. *Ann. Appl. Stat.* 3(1): 422–457. DOI: 10.1214/08-AOAS200.
- Bali, T. G., N. Cakici, and H. Levy (2008). “A model-independent measure of aggregate idiosyncratic risk”. *Journal of Empirical Finance*. 15(5): 878–896. ISSN: 0927-5398. DOI: <https://doi.org/10.1016/j.jempfin.2008.02.002>.
- Bartram, S. M., G. W. Brown, and R. M. Stulz (2018). “Why Has Idiosyncratic Risk Been Historically Low in Recent Years?” *Tech. rep.* National Bureau of Economic Research.
- Bekaert, G., R. J. Hodrick, and X. Zhang (2012). “Aggregate Idiosyncratic Volatility”. *Journal of Financial and Quantitative Analysis*. 47(6): 1155–1185. DOI: 10.1017/S0022109012000543.
- Bennett, J. A., R. W. Sias, and L. T. Starks (2003). “Greener pastures and the impact of dynamic institutional preferences”. *The Review of Financial Studies*. 16(4): 1203–1238.
- Bessembinder, H. (2000). “Tick Size, Spreads, and Liquidity: An Analysis of Nasdaq Securities Trading near Ten Dollars”. *Journal of Financial Intermediation*. 9(3): 213–239. ISSN: 1042-9573. DOI: <https://doi.org/10.1006/jfin.2000.0288>.
- Bollen, N. P., T. Smith, and R. E. Whaley (2004). “Modeling the bid/ask spread: measuring the inventory-holding premium”. *Journal of Financial Economics*. 72(1): 97–141. ISSN: 0304-405X. DOI: [https://doi.org/10.1016/S0304-405X\(03\)00169-7](https://doi.org/10.1016/S0304-405X(03)00169-7).
- Brandt, M. W., A. Brav, J. R. Graham, and A. Kumar (2010). “The Idiosyncratic Volatility Puzzle: Time Trend or Speculative Episodes?” *Review of Financial Studies*. 23(2): 863–899.
- Brown, G. and N. Kapadia (2007). “Firm-specific risk and equity market development”. *Journal of Financial Economics*. 84(2): 358–388. ISSN: 0304-405X. DOI: <https://doi.org/10.1016/j.jfineco.2006.03.003>.
- Campbell, J. Y., A. W.-C. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- Campbell, J. Y., M. Lettau, B. G. Malkiel, and Y. Xu (2001). “Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic

- Risk". *The Journal of Finance*. 56(1): 1–43. ISSN: 1540-6261. DOI: 10.1111/0022-1082.00318.
- Christie, W. G. and P. H. Schultz (1994). "Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *The Journal of Finance*. 49(5): 1813–1840.
- Chun, H., J.-W. Kim, R. Morck, and B. Yeung (2008). "Creative destruction and firm-specific performance heterogeneity". *Journal of Financial Economics*. 89(1): 109–135.
- Comin, D. and S. Mulani (2009). "A theory of growth and volatility at the aggregate and firm level". *Journal of Monetary Economics*. 56(8): 1023–1042.
- Comin, D. and T. Philippon (2005). "The Rise in Firm-Level Volatility: Causes and Consequences". *NBER Macroeconomics Annual*. 20: 167–201. DOI: 10.1086/ma.20.3585419. eprint: <https://doi.org/10.1086/ma.20.3585419>.
- Corwin, S. A. and P. Schultz (2012). "A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices". *The Journal of Finance*. 67(2): 719–760.
- Doidge, C., G. A. Karolyi, and R. M. Stulz (2017). "The U.S. listing gap". *Journal of Financial Economics*. 123(3): 464–487. ISSN: 0304-405X. DOI: <https://doi.org/10.1016/j.jfineco.2016.12.002>.
- Doucet, A. and A. M. Johansen (2011). "A tutorial on particle filtering and smoothing: fifteen years later".
- Epps, T. W. (1979). "Comovements in Stock Prices in the Very Short Run". *Journal of the American Statistical Association*. 74(366): 291–298. ISSN: 01621459.
- Fama, E. F. and K. R. French (1997). "Industry costs of equity". *Journal of Financial Economics*. 43(2): 153–193. ISSN: 0304-405X. DOI: [https://doi.org/10.1016/S0304-405X\(96\)00896-3](https://doi.org/10.1016/S0304-405X(96)00896-3).
- Fink, J., K. E. Fink, G. Grullon, and J. P. Weston (2010). "What drove the increase in idiosyncratic volatility during the internet boom?" *Journal of Financial and Quantitative Analysis*. 45(5): 1253–1278.
- Gaspar, J.-M. and M. Massa (2006). "Idiosyncratic volatility and product market competition". *The Journal of Business*. 79(6): 3125–3152.
- Hasbrouck, J. (1999). "The Dynamics of Discrete Bid and Ask Quotes". *The Journal of Finance*. 54(6): 2109–2142. ISSN: 00221082, 15406261.
- Hasbrouck, J. (2009). "Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data". *The Journal of Finance*. 64(3): 1445–1477.

- Hoberg, G. and G. Phillips (2010). "Product market synergies and competition in mergers and acquisitions: A text-based analysis". *The Review of Financial Studies*. 23(10): 3773–3811.
- Hoberg, G. and G. Phillips (2017). "Text-Based Industry Momentum". *Journal of Financial and Quantitative Analysis*. forthcoming.
- Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka (1999). "A New Estimate of Transaction Costs". *The Review of Financial Studies*. 12(5): 1113–1141.
- Lesmond, D. A., X. Pan, R. Stein, and Y. Zhao (2018). "What Drives the Trend and Behavior in Aggregate (Idiosyncratic) Variance? Follow the Bid-Ask Bounce". *SSRN Working Paper*.
- Pontiff, J. and P. Irvine (2009). "Idiosyncratic return volatility, cash flows, and product market competition". *Review of Financial Studies*. 22(3): 1149–1177.
- Roll, R. (1984). "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market". *The Journal of Finance*. 39(4): 1127–1139.
- Vogelsang, T. (1998). "Trend Function Hypothesis Testing in the Presence of Serial Correlation". *Econometrica*. 66(1): 123–148.
- Wei, S. X. and C. Zhang (2006). "Why did individual stocks become more volatile?" *The Journal of Business*. 79(1): 259–292.
- Xu, Y. and B. G. Malkiel (2003). "Investigating the Behavior of Idiosyncratic Volatility". *The Journal of Business*. 76(4): 613–645. ISSN: 00219398, 15375374.
- Zhang, M. Y., J. R. Russell, and R. S. Tsay (2008). "Determinants of bid and ask quotes and implications for the cost of trading". *Journal of Empirical Finance*. 15(4): 656–678. ISSN: 0927-5398. DOI: <https://doi.org/10.1016/j.jempfin.2007.12.003>.

Appendix

A Derivation of Equation (8)

In the derivation we suppress the time subscripts for brevity. We first rewrite the IV, defined in Equation (5), as

$$\sigma_\eta^2 = \sum_i w_i \sum_{j \in i} w_{ji} \text{Var}(R_{ji} - R_i) \quad (\text{A1})$$

$$= \sum_i w_i \sum_{j \in i} w_{ji} (\text{Var}(R_{ji}) + \text{Var}(R_i) - 2 \text{Cov}(R_{ji}, R_i)). \quad (\text{A2})$$

Next, we note that

$$\sum_i w_i \sum_{j \in i} w_{ji} \text{Var}(R_i) = \sum_i w_i \text{Var}(R_i) \quad (\text{A3})$$

$$\sum_i w_i \sum_{j \in i} w_{ji} \text{Cov}(R_{ji}, R_i) = \sum_i w_i \text{Cov} \left(\sum_{j \in i} w_{ji} R_{ji}, R_i \right) \quad (\text{A4})$$

$$= \sum_i w_i \text{Cov}(R_i, R_i) = \sum_i w_i \text{Var}(R_i). \quad (\text{A5})$$

Plugging in the above in Equation (A2) gives the desired result.

B Derivation of Equation (9)

In the derivation we suppress the time subscripts for brevity. Equation (8) shows that we need to only compute average firm and industry variances. Let \bar{w} denote the vector of weights within an industry. Direct computation

gives

$$\sigma_\eta^2 = \sum_i w_i \sum_{j \in i} w_{ji} \text{Var}(R_{ji}) - \sum_i w_i \text{Var}(R_i) \quad (\text{B6})$$

$$\sum_i w_i \sum_{j \in i} w_{ji} \text{Var}(R_{ji}) = \sum_i w_i \sum_{j \in i} w_{ji} \bar{\sigma}_i^2 = \sum_i w_i \bar{\sigma}_i^2 \quad (\text{B7})$$

$$\sum_i w_i \text{Var}(R_i) = \sum_i w_i \bar{\mathbf{w}}^\top \bar{\sigma}_i^2 ((1 - \rho)I + \rho \mathbf{1}\mathbf{1}^\top) \bar{\mathbf{w}} \quad (\text{B8})$$

$$= \sum_i w_i \bar{\sigma}_i^2 \left((1 - \rho) \sum_j \bar{w}_j^2 + \rho \right). \quad (\text{B9})$$

Combining the average variance of firms and industries verifies Equation (9).

C Filtering procedure

To estimate our nonlinear state-space model outlined in Equations (13) to (18), we implement a particle filter.³⁰ We distinguish three cases: when only closing prices are available (CP), when bid and ask quotes are available (BA), and when price information is unavailable (NA). The state vector f is four-dimensional, $f = \{s^*, \log(\sigma), c, \chi\}$. We first sample $\log(\sigma_0)$ from a standard Gaussian distribution with a mean equal to a logarithm of Roll's estimate based on the first 22 return observation as measured by CRSP return series. We bound Roll's standard deviation by 0.001 from below. For χ_0 , we use a Gaussian distribution with standard deviation of 0.2. We set the mean to $\log(\log(A_1/B_1)/4)$ in the BA case. In the CP case, we use the log-spread implied by the model of Roll, equal to $\log(\sqrt{-\text{Cov}(r_t, r_{t-1})})$, and estimate the autocorrelation based on the first 22 returns. In case the autocorrelation is non-negative, we use $\log(\log(\min\{(S_1 + d_1)/(S_1 - d_1), 5\})/4)$ instead. We draw s_0^* from a Gaussian distribution with standard deviation equal to $2e^{\chi_0}$, centered either around the log of the price on the first day (CP), or around the average of the log-bid and log-ask (BA).

³⁰We refer a reader unfamiliar with particle filtering to the introductory material of Doucet and Johansen (2011).

We denote observations at time t by Y_t , which consists either of the bid-ask pair or the closing price. The filtering procedure sequentially approximates $p(f_{1:t}|Y_{1:t}) \propto p(Y_t|f_t)p(f_t|f_{t-1})p(f_{1:t-1}|Y_{1:t-1})$ by a distribution over a discrete set of particles. The transition densities are specified by Equations (15) to (18), and $p(f_{1:t-1}|Y_{1:t-1})$ is approximated in the previous step of the filter. In the BA case, when Y_t consists of the observed bid and ask quote B_t and A_t , Equation (14) implies

$$p(Y_t|f_t) \propto \mathbb{1}_{\{\log(A_t - d_t) \leq s_t^* + C_t \leq \log(A_t) \wedge \log(B_t) \leq s_t^* - C_t \leq \log(B_t + d_t)\}}. \quad (C10)$$

Therefore, the observation equation only imposes constraints on the price-spread pair (rectangular for $\{s^*, C\}$). Because $c = \log(C)$ as well as $\log(S^*F + X)$ are Gaussian, conditionally on current volatility and previous states, the target distribution is bivariate Gaussian constrained to a set with nonlinear boundaries. Due to these restrictions, it is desirable to avoid sampling outside of the feasible set of state values. Next, we denote the normal distribution truncated to the interval $[l, u]$ by $tN(\mu, \sigma^2, l, u)$ and by $tN(\mu, \sigma^2, I)$ when truncated to the set I . Further denoting the i -th particle by the superscript (i) , we sample the states as follows to ensure that the draws are within the desired region.

1. We draw log-volatilities from the Gaussian transition equation (16).
2. We sample $c_t^{(i)}$ from $tN(\xi_{t-1}^{(i)}, \gamma_u^2 + \gamma_\chi^2, k_l, k_u)$. The bounds of the truncation region are equal to

$$k_l = \log(\max((\log(A_t - d_t) - \log(B_t + d_t))/2, 0)), \quad (C11)$$

$$k_u = \log((\log(A_t) - \log(B_t))/2). \quad (C12)$$

3. Conditional on draws $\sigma^{(i)}$ and $c_t^{(i)}$, we sample $\log(S^*F + X)^{(i)}$ from $tN(s_{t-1}^{(i)}, (\sigma^2)^{(i)}, j_l, j_u)$, with

$$\tilde{j}_l = \max(\log(A_t - d_t) - C_t^{(i)}, C_t^{(i)} + \log(B_t)), \quad (C13)$$

$$\tilde{j}_u = \min(\log(A_t) - C_t^{(i)}, C_t^{(i)} + \log(B_t + d_t)), \quad (C14)$$

$$j_l = \log(\exp(\tilde{j}_l)F_t + X_t), \quad (C15)$$

$$j_u = \log(\exp(\tilde{j}_u)F_t + X_t). \quad (C16)$$

The \tilde{j} 's correspond to bounds on the log-price, the j 's account for dividend payments.

4. Conditional on $c_t^{(i)}$, we sample $\chi_t^{(i)}$ from $N(\mu_\chi, \sigma_\chi^2)$, where

$$\mu_\chi = \bar{\chi} + \varphi_\chi(\chi_t - \bar{\chi}) + \frac{\gamma_\chi^2}{\gamma_c^2 + \gamma_\chi^2}(c_t^{(i)} - (\bar{\chi} + \varphi_\chi(\chi_t - \bar{\chi}))), \quad (\text{C17})$$

$$\sigma_\chi = \frac{\gamma_c^2 \gamma_\chi^2}{\gamma_c^2 + \gamma_\chi^2}. \quad (\text{C18})$$

In the CP case, we observe only the closing price, which can be either of ask, so that $s_t^* + C_t \in [S_t - d_t, S_t]$, or bid which implies $s_t^* - C_t \in [S_t, S_t + d_t]$. Thus,

$$p(Y_t | f_t) \propto 0.5 \mathbb{1}_{\{\log(P_t - d_t) \leq s_t^* + C_t \leq \log(P_t)\}} + 0.5 \mathbb{1}_{\{\log(P_t) \leq s_t^* - C_t \leq \log(P_t + d_t)\}}. \quad (\text{C19})$$

For the CP case, the following sampling scheme guarantees that we draw in the feasible region.

1. We draw log-volatilities from the Gaussian transition equation (16).
2. We sample $\{c_t^{(i)}, \chi_t^{(i)}\}$ from their transition equations (17)-(18).
3. Conditional on draws $\sigma^{(i)}$ and $c_t^{(i)}$, we sample $\log(S_t^* F_t + X_t)^{(i)}$ from $tN(s_{t-1}^{(i)}, \sigma^{2(i)}, I_t)$. The truncation region is a union of two intervals, $I_t = [m_l, m_u] \cup [n_l, n_u]$, corresponding to the bid and ask cases. The interval boundaries for the log-price are defined as

$$\tilde{m}_l = \log(S_t) + C_t, \quad \tilde{m}_u = \log(S_t + d_t) + C_t, \quad (\text{C20})$$

$$\tilde{n}_l = \log(S_t - d_t) - C_t, \quad \tilde{n}_u = \log(S_t) - C_t, \quad (\text{C21})$$

and the boundaries for $\log(S_t^* F_t + X_t)^{(i)}$ are obtained analogously to Equation (C15).

When an observation is missing (NA), we directly sample from the transition equations, with zero incremental weights. After each step, we resample the particles if the effective (relative) sample size drops below one half. ³¹

³¹We use binomial resampling. We also experimented with systematic and residual resampling, but the choice of resampling method has a negligible impact on variance estimates. Therefore, we opt for the simplest scheme.

In addition, we make the number of particles N_t time-varying, increasing their count in situations when future observations are weakly informative about the states. We let the number of particles vary as $N_t = M_t \bar{N}$, where the baseline particle count \bar{N} is set to 500, and the dynamic multiplier is a smoothed version of raw multipliers defined below,

$$M_t = \max(\tilde{M}_{\max(t-99,1):t}). \quad (\text{C22})$$

The computation of raw multiplier \tilde{M} differs for the BA and CP case. In the BA case, we first flag BA observations as informative if either the bid or the ask changes ($INFT_t = 1$). Then, we compute its forward-looking moving average, $\overline{INFT}_t = \frac{1}{100} \sum_{i=0}^{99} INFT_{t+i}$, and define

$$\tilde{M}_t^{BA} = \max\left(\min\left(\left\lceil \overline{INFT}_t^{-1} \right\rceil, 10\right), 1\right), \quad (\text{C23})$$

having more particles if price changes are less frequent in future periods. For the CP case, we first compute

$$Z_t = \max\left(\frac{20}{S_t} \frac{d_t}{1/8}, 1\right), \quad (\text{C24})$$

then we take its forward-looking average,

$$\bar{Z}_t = \frac{1}{100} \sum_{i=0}^{99} Z_{t+i}, \quad (\text{C25})$$

and finally define

$$\tilde{M}_t^{CP} = \min(\lceil Z_t \rceil, 10). \quad (\text{C26})$$

This criterion increases the number of particles in cases the price is low relative to the prevailing tick size. By taking the rolling maximum in Equation (C22), we avoid frequent oscillation in the particle count when the BA and CP cases switch due to lack of trading activity and imply a different raw multiple.

The next ingredient in the implementation of the filter is specification of the tick-size d_t . We extract the values for each stock individually. For each date we take the set of all quotes and trading prices (not midpoints) that occurred up to date, and find the smallest difference among the prices in the set. For the first 22 observations, we use the window of first 22

days of prices. An advantage of this approach is its simplicity and that it is able to capture aspects such as gradual implementation of quote decimalization, or different rules of exchanges applying to cross-listed stocks. One disadvantage of this procedure is that the tick reduction is permanent. In particular, since for stocks priced below 1\$ the tick sizes are lower, recovery of the price above this threshold will result in an underestimation of d_t . The same issue may arise in case of quote errors, or unusual quotes (ticks) on NASDAQ, where the size of price increments was restricted by customs rather than formal rules. If the tick size is too low, the observed and the latent quotes almost coincide and the filtered price is close to the quote midpoint, due to the assumption of a symmetric spread.

We keep the parameters of the transition dynamics constant and identical for all stocks, which avoids the computational burden that would stem from a formal estimation. We set the mean-reversion parameters φ_σ and φ_χ to 0.999, implying a (prior) half-life of 623 days.³² We fix the reversion levels to $\log(\bar{\sigma} = \log(0.5/252))$ and $\bar{\chi} = \log(\log(1.25)) - \log(2)$. However, the exact choice of the constant is not crucial due to relatively long half-life of the processes. For the volatility in the mean spread, we choose $\gamma_\chi = 0.02$, motivated by slow variation in long-term spread levels. To capture the spread oscillation around its local level, we let $\gamma_c = 0.5$. This choice becomes vital to capture the spread volatility following the switch to inside quotes for NASDAQ stocks in 1980. Finally, we set the volatility-of-volatility to $\gamma_\sigma = 0.1$. This choice allows for fast changes in volatility. A two-standard deviation increase over ten-day horizon corresponds to drop of volatility by -53% or an increase by 88%.

Before applying the filtering procedure outlined above, we eliminate outliers from the data, which often stem from data errors (unrepresentative quotes). Panels A and B in Figure C1 show two such observations, one corresponding to a closing price and the other to a bid-ask midpoint.

We identify suspicious quotes in the data by combining multiple criteria. First, we flag as a quote error observations with extreme or unusually wide

³²We also experimented with a unit root specification instead of mean-reverting process. Overall, the estimates are barely affected by this choice, but the mean-reverting specification is more robust to long, uninformative periods.

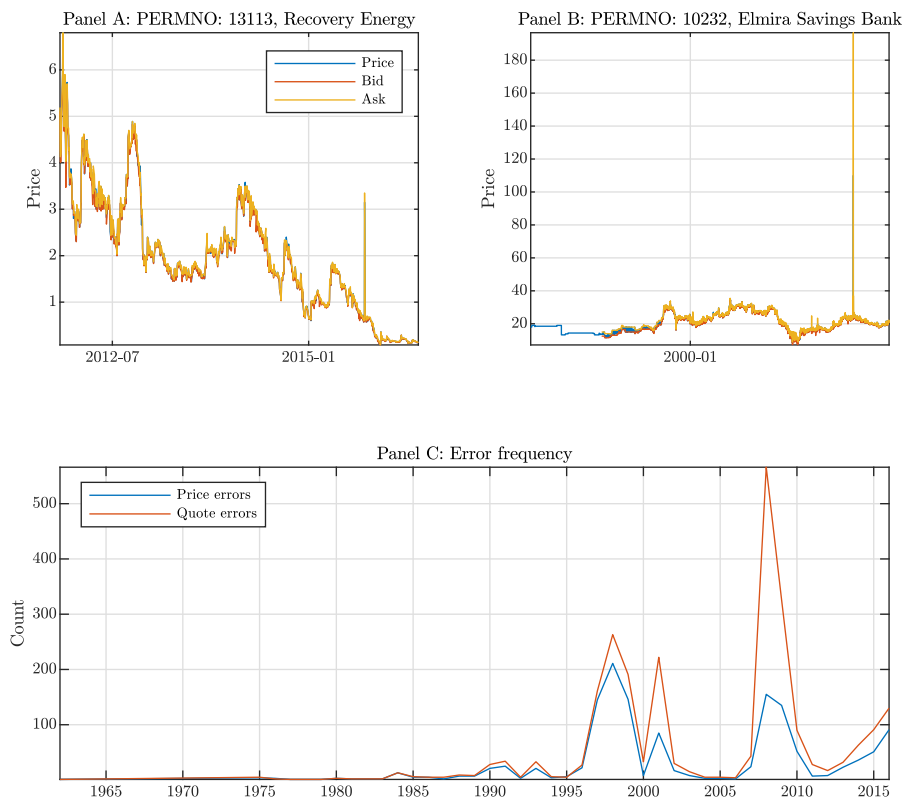


Figure C1: CRSP problematic quotes

Description: Panel A: Example of a stock with outlier price observation, corresponding to closing price. Panel B: Example of a stock with outlier price observation, corresponding to quote midpoint. Panel C: Time series of outlier occurrences.

Interpretation: CRSP data contain numerous outliers, some of which correspond to data errors or unrepresentative quotes. Data cleaning is therefore important.

bid-ask spread, identified as

$$C_{ES} = \frac{A}{B} > 5, \quad (C27)$$

$$C_{US} = \left| \text{med} \left(\log \left(\frac{A}{B} \right), 9 \right) - \log \left(\frac{A}{B} \right) \right| > 1, \quad (C28)$$

$$C_{Q1} = C_{ES} \vee C_{US}. \quad (C29)$$

Then, we add a criterion to capture extreme quotes, which still possibly exhibit a reasonable (or less extreme) spread. First, we check whether either the spread is wide (in absolute or relative terms) or the recorded price is outside of prevailing quotes. Formally,

$$C_{REL} = \max \left(1, \frac{(A-1/8)}{(B+1/8)} \right) - 1 > 1, \quad (C30)$$

$$C_{ABS} = \left(\max \left(1, \frac{(A-1/8)}{(B+1/8)} \right) - 1 > \frac{1}{2} \right) \wedge (A-B > 3), \quad (C31)$$

$$C_{OUT} = (P > A) \vee (P < B). \quad (C32)$$

and combine them to $C_{QCS} = C_{REL} \vee C_{ABS} \vee C_{OUT}$. Then, we compare the log-bid and log-ask, adjusted for dividends and stock splits, with their moving medians over a window of 9 days. We check the same criterion on bid (ask) level as well, and consider the quote to be problematic if both the level and logarithmic criteria hold. We combine the spread criterion C_{QCS} with the time-series criterion C_{QTS} to obtain the final criterion for erroneous quotes C_Q ,

$$C_{QTS} = \left(\left(|b - \text{med}(b, 9)| > \frac{1}{2} \right) \wedge \left(|B - \text{med}(B, 9)| > \frac{1}{2} \right) \right) \vee \left(\left(|a - \text{med}(a, 9)| > \frac{1}{2} \right) \wedge \left(|A - \text{med}(A, 9)| > \frac{1}{2} \right) \right), \quad (C33)$$

$$C_{Q2} = C_{QTS} \wedge C_{QCS}, \quad (C34)$$

$$C_Q = C_{Q1} \vee C_{Q2}. \quad (C35)$$

For the price series, we label the observation as suspicious if the log-price is far away from its moving median over a window of five observations. To avoid false positives for low-priced stocks, we also require the same criterion to hold for price levels. Formally, we let

$$C_{PREL} = |p - \text{med}(p, 5)| > 1, \quad C_{PABS} = |P - \text{med}(P, 5)| > 1. \quad (C36)$$

and $C_{PTS} = C_{PREL} \wedge C_{PABS}$. In addition, we also treat all prices computed as quote midpoint, when the quotes are erroneous (C_Q holds) as an error, $C_P: C_{PTS} \vee (C_Q \wedge \text{Midpoint})$. In both cases we keep the last observation in the sample, whether the criteria described above hold or not, so that we do not exclude pre-delisting information. Table C1 provides examples of detected errors by our cleaning procedure.³³ Panel C in Figure C1 shows that most of the outliers are detected around the dot-com bubble, the financial crisis, and in the most recent period. The surge in the former two periods likely reflects a combination of low liquidity and false positives. In shallow markets, a huge price change may occur due to the price impact of a block trade. Extreme quotes often occur due to the unwillingness of the market maker to trade, motivating them to post extreme quotes. The increase in the recent period is more puzzling, perhaps indicating that the unrepresentative quotes are only gradually assessed and eliminated by CRSP

Using the criteria above we distinguish several cases. If both C_P and C_Q hold (or quotes are unavailable) then we treat the observation as missing, i.e., the NA case. If only C_Q holds, we eliminate the quotes and use the closing price only (CP). When the quotes are error-free, but the price is not, the effect on our filtering procedure is limited, as in such cases we use the quotes. Still, the price error might cause problems for the initialization of the filter, so we convert an erroneous price to its previous value and recompute the resulting returns.³⁴

From the particle filter we compute two sets of estimates, a filtered

³³In most cases, the presence of a quote “error” is clear. For example, in the second line of Table C1, the observation is flagged as problematic because the ask/bid is too large. With a wide bid-ask spread, the discrepancy between midpoints under assumptions of a symmetric spread in levels and logs becomes large. As a consequence, the resulting filtered returns might be large if the adjacent observations have either narrower spread or do not have available quotes. As another example, the observation in the sixth block is possibly a false positive, flagged because of its relatively low bid compared to neighboring observations, and the fact that the price is outside of the quotes. As a consequence, the filtering procedure uses the price instead of the quotes.

³⁴Furthermore, price information might be missing for other reasons than our exclusion of outliers, e.g., due to suspension from trading. We treat those missing values analogously, i.e., as the NA case, unless there are missing values for more than 22 consecutive observation. In such cases, the variance of the filter would be too large. Instead, we split the full data range of the stock into connected segments, where no such gaps occur, and estimates the states on each segment separately.

ID	Date	PRC	BID	ASK	BIDLO	ASKHI	RET	Mid
10001	03-Aug-2009	8.050	8.050	8.150	7.750	8.340	-0.012	false
10001	04-Aug-2009	8.600	7.750	999.990	8.400	8.600	0.068	false
10001	05-Aug-2009	8.578	8.350	8.600	7.960	8.590	-0.003	false
10042	05-Aug-1999	0.188	-	-	0.156	0.188	0.200	false
10042	06-Aug-1999	0.172	0.031	0.313	0.031	0.313	-0.083	true
10042	09-Aug-1999	0.156	-	-	0.156	0.188	-0.091	false
10100	11-Apr-2008	1.050	0.820	1.280	0.820	1.280	-0.014	true
10100	14-Apr-2008	2.595	0.710	4.480	0.710	4.480	1.471	true
10100	15-Apr-2008	1.035	0.820	1.250	0.820	1.250	-0.601	true
10100	09-May-2008	1.025	0.850	1.200	0.850	1.200	0.000	true
10100	12-May-2008	0.850	0.350	1.780	0.850	0.850	-0.171	false
10100	13-May-2008	1.065	0.850	1.280	0.850	1.280	0.253	true
10100	26-Sep-2008	0.775	0.670	0.880	0.670	0.880	-0.119	true
10100	29-Sep-2008	0.880	0.250	1.250	0.880	0.880	0.136	false
10100	30-Sep-2008	0.670	0.790	0.880	0.670	0.790	-0.239	false
10100	18-Dec-2008	0.520	0.490	0.550	0.490	0.550	-0.096	true
10100	19-Dec-2008	0.310	0.390	0.440	0.310	0.530	-0.404	false
10100	22-Dec-2008	0.310	0.310	0.820	0.310	0.310	0.000	false
10100	13-Jan-2009	0.600	0.710	0.790	0.600	0.680	-0.143	false
10100	14-Jan-2009	1.390	0.040	2.740	0.040	2.740	1.317	true
10100	15-Jan-2009	0.725	0.680	0.770	0.680	0.770	-0.478	true
10205	26-Sep-2008	12.000	11.900	12.010	12.000	12.970	-0.016	false
10205	29-Sep-2008	10.800	6.000	12.100	9.050	12.100	-0.100	false
10205	30-Sep-2008	10.650	10.650	10.660	10.550	12.440	-0.014	false
10232	09-Dec-2013	25.200	23.350	29.750	23.201	25.200	0.019	false
10232	10-Dec-2013	110.035	23.400	196.670	23.400	196.670	3.366	true
10232	11-Dec-2013	24.010	23.520	25.000	23.310	25.250	-0.782	false
10256	28-Jun-2001	0.600	0.530	0.600	0.550	0.600	0.091	false
10256	29-Jun-2001	0.550	0.110	3.000	0.550	0.570	-0.083	false
10256	02-Jul-2001	0.550	0.500	0.550	0.490	0.550	0.000	false

Table C1: Examples of flagged observations by error detection procedure.

Description: Table of first ten detected errors. Each block shows period $t - 1$ to $t + 1$ for an error detected for time t . ID column contains PERMNO, i.e., the CRSP security identifier. PRC is the price series, i.e., an absolute value of price series from CRSP database. BID and ASK are the closing or inside quotes, depending on exchange under consideration. BIDLO and ASKHI are the closing (inside) quotes when the trading price is not available (Midpoint), and the daily low and high price otherwise. Mid indicates whether the price corresponds to quote midpoint.

Interpretation: The error identification procedure successfully identifies many problematic quotes that often generate huge artificial returns.

series and a smoothed series. The latter is obtained from a fixed-lag approximation using $L = 100$. While other smoothing methods might be preferable, we opt for the fixed-lag approximation for its simplicity and low computational cost, which is of practical relevance given that we apply the filter for approximately 24,000 stocks.